UNIVERSITAT DE BARCELONA

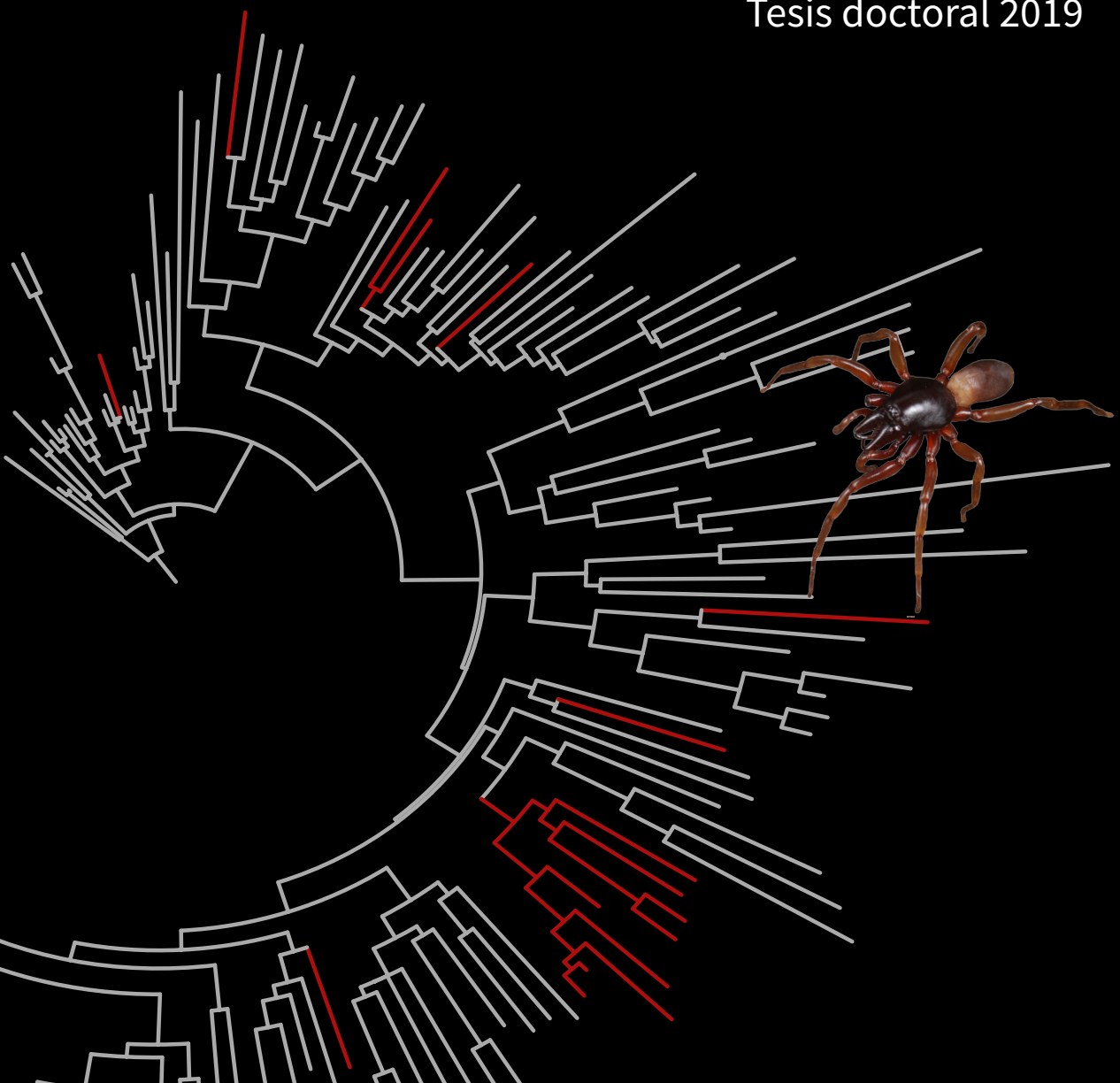
Genómica de la adaptación en artrópodos: estudio del sistema quimiosensorial y de la radiación del género *Dysdera* (Araneae) en Canarias

Joel Vizueta Moraga

Genómica de la adaptación en artrópodos: estudio del sistema quimiosensorial y de la radiación del género *Dysdera* (Araneae) en Canarias

Joel Vizueta Moraga

Tesis doctoral 2019



UNIVERSITAT DE BARCELONA

Genómica de la adaptación en artrópodos: estudio del sistema quimiosensorial y de la radiación del género *Dysdera* (Araneae) en Canarias

Memoria presentada por **Joel Vizueta Moraga**
para optar al Grado de Doctor por la Universidad de Barcelona.

Departamento de Genética, Microbiología y Estadística

El autor de la tesis
Joel Vizueta Moraga

El tutor y codirector de la tesis
Dr. Julio Rozas Liras
Catedrático de Genética
Departamento de Genética,
Microbiología y Estadística
Facultad de Biología
Universidad de Barcelona

El codirector de la tesis
Dr. Alejandro Sánchez Gracia
Profesor asociado de Genética
Departamento de Genética,
Microbiología y Estadística
Facultad de Biología
Universidad de Barcelona

Barcelona, Septiembre de 2019

"Tener una idea es lo más fácil del mundo. Todo el mundo tiene ideas. Pero tienes que tomar esa idea y convertirla en algo a lo que la gente responderá, eso es difícil"

Stanley Martin Lieber

Gracias a todos los que habéis hecho posible esta tesis, compañeros, amigos, familia...

Y en especial a Paula, Skye, a mis padres y mi hermano Álvaro

Abstract

During their evolutionary history, arthropods have diversified adapting to different habitats, including several independent colonizations of land, and sometimes implicating rapid radiations coupled with dietary specializations. Although the chemosensory system likely played a critical role in many of these adaptations, the origin and evolution of the gene families that mediate chemoperception in arthropods is still discussed in some important aspects. The main objective of this thesis is to gain insights into the molecular evolution of chelicerate diversification and, specifically, to determine the role of natural selection in this process. On one hand, we studied the evolution of the chemosensory gene families in chelicerates using comparative transcriptome and genome analyses. We first developed a bioinformatic pipeline (BITACORA) for the identification and annotation of gene families in genome assemblies. Using this tool, we identified members of two of the major arthropod chemoreceptor gene families (GRs and IRs) in chelicerates, being some of them expressed in the chemosensory appendages of the spider *Dysdera silvatica*, which supports its role in chemoperception. These families evolved under a dynamic gene birth and death model influenced by episodic bursts of gene duplication yielding lineage-specific expansions. Noticeably, we characterized in chelicerates a gene family distantly related to insect OBPs, suggesting a more ancient origin of these soluble carriers than previously thought, and a new gene family encoding small globular secreted proteins, which is a good chemosensory gene family candidate. In addition, we discuss the absence of the CSP family in chelicerates, and the putative role of NPC2 members in chemoperception. On the other hand, we studied the radiation of the spider genus *Dysdera* in the Canary Islands, where species diversification occurs concomitant with repeated events of trophic specialization. We identified a number of genetic changes likely associated with this convergent adaptation, including some related to heavy metal detoxification and homeostasis, metabolism of important nutrients and venom toxins. We uncovered the specific molecular substrates of these changes at different hierarchical levels, including same genes, gene functions or amino acid positions, some of them promoted by positive selection. Globally, our results increase the knowledge about the molecular basis of adaptation and provide new insights into the predictability of evolution.

Índice

Introducción 1

1 Genómica evolutiva 3
- 1.1 Evolución molecular 3
- 1.2 Selección natural a nivel molecular 4
- 1.3 Radiaciones adaptativas 6
- 1.4 Convergencia y predictibilidad en la evolución 7

2 El sistema quimiosensorial 9
- 2.1 El sistema quimiosensorial y su papel en la adaptación 9
- 2.2 El sistema quimiosensorial periférico en artrópodos 11
- 2.3 El sistema quimiosensorial en artrópodos: familias multigénicas 13
  - 2.3.1 Quimiorreceptores 13
  - 2.3.2 Proteínas de unión a ligando 17
- 2.4 Origen y evolución de las familias multigénicas 19

3 Organismos de estudio: Quelicerados 21
- 3.1 El género *Dysdera* 22

Objetivos 27

Informe de los directores 31

Capítulos 35

1 BITACORA: A comprehensive tool for the identification and annotation of gene families in genome assemblies 37

2 Evolution of chemosensory gene families in arthropods: Insight from the first inclusive comparative transcriptome analysis across spider appendages 73

3 Comparative genomics reveals thousands of novel chemosensory genes and massive changes in chemoreceptor repertories across chelicerates 111

4 Chance and predictability in evolution: the genomic basis of convergent dietary specializations in an adaptive radiation 143

Discusión 185

1 Desarrollo e implementación de nuevos métodos para el estudio de familias multigénicas en ensamblajes genómicos 188

2 Origen y evolución de las familias multigénicas del sistema quimiosensorial en artrópodos 190

2.1 Quimiorreceptores en artrópodos 190

2.2 Proteínas solubles secretadas 194

3 Determinantes genómicos de la especialización trófica convergente en *Dysdera* 196

Conclusiones 201

Bibliografía 209

Anexo 223

A Comparative analysis of tissue-specific transcriptomes in the funnel-web spider *Macrothele calpeiana* (Araneae, Hexathelidae) 225

B Financiación 251

Introducción

1 Genómica evolutiva

1.1 Evolución molecular

El concepto actual de evolución biológica establece que todas las especies existentes han derivado a partir de un ancestro común por la acumulación de cambios graduales de forma independiente en los distintos linajes ocasionando su diferenciación. En "El origen de las especies", Charles Darwin[1] introdujo el concepto de evolución por selección natural donde formula que la evolución se produce por la aparición al azar de variantes que pueden conferir ventajas adaptativas a los individuos de la población, incrementando su probabilidad de supervivencia y reproducción, en respuesta a circunstancias ambientales cambiantes. Actualmente, la base de la genética evolutiva postula que los principales mecanismos que explican los patrones de variación genética entre poblaciones y especies son: la mutación, la deriva genética, la migración, la recombinación y la selección natural. Estas fuerzas pueden actuar de forma paralela, viéndose influenciadas por factores demográficos, y son las que explican los patrones evolutivos observados en las poblaciones y especies[2].

El modelo más aceptado actualmente que explica de forma global los patrones de variación molecular, y único en alcanzar el nivel de teoría, es la teoría neutralista de evolución molecular. Esta teoría fue propuesta independientemente por Motoo Kimura[3] y Jack King y Thomas Jukes[4], y sostiene que, a nivel molecular, la gran mayoría de la variación genética no está moldeada por la selección natural positiva (o selección *darwiniana*), sino por la segregación y/o fijación al azar de sustituciones neutras o casi-neutras (Figura 1)[5,6]. No obstante, habrá una pequeña fracción de las variantes que estarán afectadas por la selección natural debido a su impacto en la eficacia biológica o aptitud de los organismos (conocido como *fitness*). Las mutaciones deletéreas, es decir, aquellas que afecten negativamente a la eficacia biológica de un organismo, estarán influenciadas por la selección negativa o purificadora, provocando que los individuos portadores tengan una menor probabilidad de reproducirse y, por tanto, de pasar dicha variante a su descendencia, reduciendo así la frecuencia de estas mutaciones en la población pudiendo llegar a

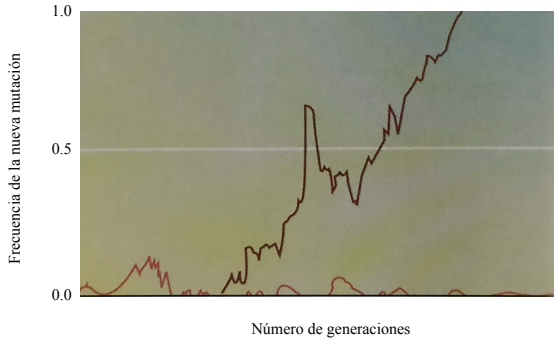


Figura 1. Aparición, pérdida y eventual fijación de nuevas mutaciones en una población de pequeño tamaño poblacional. La teoría neutralista establece que la mayoría de mutaciones son neutras y evolucionan por deriva genética y, por tanto, su destino depende del azar. La figura muestra el origen de 10 mutaciones, 9 de las cuales se perdieron tras aumentar ligeramente de frecuencia (en color rojo), y una llegó a fijarse en la población (de color marrón). Adaptado de Crow y Kimura[7].

ser eliminadas. Sin embargo, las mutaciones beneficiosas, las cuales incrementan la eficacia biológica del individuo, evolucionarán por selección positiva *darwiniana*, resultando en una mayor probabilidad de supervivencia y reproducción del organismo, y el consecuente aumento de su frecuencia en la población, pudiendo llegar a fijarse[8]. En consecuencia, esta pequeña fracción de variantes que estarán afectadas por la selección natural tendrá una alta probabilidad de perderse o fijarse y, por tanto, tendríamos poca probabilidad de verlas segregando en las poblaciones. De esta forma, la selección natural es la fuerza predominante que explica las adaptaciones de las especies al medio.

1.2 Selección natural a nivel molecular

La detección de la selección natural es fundamental en genómica evolutiva dado que nos permite comprender los procesos adaptativos. Existen distintas aproximaciones para la detección de selección, entre las cuales destacan las proporcionadas por los tests de neutralismo. Estos tests usan la teoría neutralista como hipótesis nula, partiendo de la premisa de que todas las mutaciones son neutras (no tienen efecto en la eficacia biológica), y permiten determinar de forma estadística desviaciones del modelo nulo (rechazo de H_0). Tales desviaciones pueden ser consecuencia de la acción de la selección natural, que típicamente afectaría a un bajo número de genes o regiones genómicas, pero también reflejo de eventos demográficos, como puede ser un cuello de botella, el cual afectaría a una parte importante, o incluso a todo el genoma.

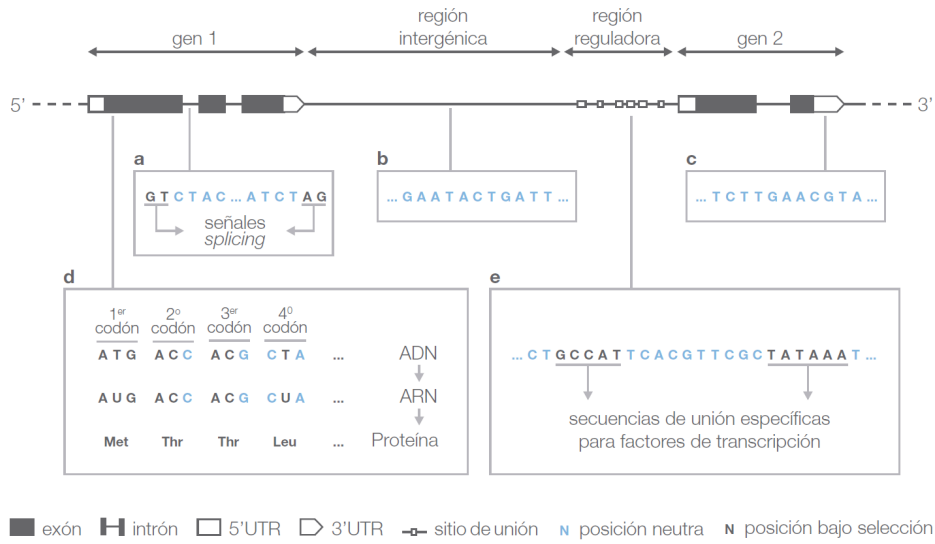


Figura 2. Ubicación y efecto de nuevas mutaciones en una región genómica. En azul se indican aquellas posiciones dónde las mutaciones no supondrían un cambio fenotípico, de forma general, y por tanto serían selectivamente neutras. Tal es el caso de la mayoría de mutaciones ubicadas en intrones, regiones intergénicas, regiones no traducidas (UTR) y posiciones sinónimas en la región codificadora (representadas como **a**, **b**, **c** y **d**, respectivamente). Por el contrario, en negro se indican las posiciones que supondrían un cambio en la estructura primaria (como por ejemplo en **a**, mutaciones en señales de *splicing*, o en **d**, cambios no sinónimos en la región codificadora) o en la expresión (**e**, sustituciones en la región reguladora) de la proteína, y, por tanto, estarían bajo el efecto de la selección natural, dependiendo su dirección e intensidad según el efecto de la mutación en la eficacia biológica del organismo. Fuente: Calvo-Martín[9].

Una de las aproximaciones más populares para detectar tales episodios de selección en regiones codificadoras a nivel inter-específico (utilizando datos de divergencia), es el análisis del ratio entre la tasa de sustituciones no sinónimas (d_N) y sinónimas (d_S), denominado ω ($\omega = d_N/d_S$)[10]. Las sustituciones no sinónimas son aquellas que implican un cambio de aminoácido en la proteína codificante, mientras que las sinónimas codificarían para el mismo aminoácido posibilitado por la redundancia del código genético (Figura 2). El valor de ω esperado bajo un modelo estrictamente neutro, en ausencia de constricciones selectivas, sería de 1. Por consiguiente, valores significativamente inferiores a 1 serían debidos a un menor número de sustituciones no sinónimas respecto a lo esperado (asumiendo que todas las mutaciones sinónimas son neutras) y se explicarían por la actuación de la selección purificadora en la región de estudio. Por el contrario, valores significativos de ω superiores a 1 indicarían una mayor tasa de fijación de sustituciones no sinónimas que las esperadas al azar, sugiriendo la acción de la selección positiva. Sin embargo, la selección positiva podría actuar únicamente en una pequeña región genómica, viéndose dificultada su detección en estudios de $d_N > d_S$ tanto a

lo largo de todo el genoma como de un gen. Por ejemplo, en proteínas asociadas al sistema quimiosensorial en insectos, se ha detectado la huella de la selección positiva en la región implicada en la unión con sus sustratos, que comprende tan solo una pequeña fracción de la proteína[11,12]. Para poder detectar estos y otros casos específicos, se han desarrollado varios métodos probabilísticos basados en modelos de codones que aplican distintas aproximaciones, como es el ajuste de un número determinado de clases de ω para la región en cuestión, o como la estimación de ω para cada codón del gen o rama del árbol filogenético[13–15].

La huella de la selección positiva puede detectarse a nivel inter-específico (analizando sustituciones que se presuponen fijadas entre distintas especies), como en el caso previamente mencionado mediante el estudio de ω, pero también a nivel intra-específico, en este caso afectando a sucesos que han ocurrido en un periodo más reciente de tiempo (analizando las variantes presentes entre individuos de una misma población), mediante tests de neutralismo como, por ejemplo, la *D* de Tajima[16], o incluso utilizando ambos tipos de datos (polimorfismo y divergencia) como en los tests HKA y MK[17,18]. En términos generales, la identificación de la huella impresa por la selección a nivel genómico puede revelar la función adaptativa de elementos génicos como pueden ser genes, regiones reguladoras, etc., así como determinar las variantes concretas que explican los procesos adaptativos.

1.3 Radiaciones adaptativas

El conocimiento de los mecanismos evolutivos implicados en los procesos de adaptación y diversificación de las especies es necesario para el correcto manejo y conservación de la biodiversidad[19]. La biología evolutiva provee el marco conceptual para identificar el papel de la selección natural y otros mecanismos clave implicados en la diversificación a través del estudio de la variación molecular, tanto a nivel inter- como intra-específico.

La especiación, en general, es un proceso gradual que requiere la acción de distintas fuerzas evolutivas durante largos periodos de tiempo. Sin embargo, este proceso también puede ocurrir de forma rápida bajo condiciones ambientales y ecológicas inestables o cambiantes, como puede ser la colonización de un nuevo medio (denominado como radiaciones adaptativas). De hecho, las islas oceánicas son un buen ejemplo para el estudio de este proceso dado que su biota procede de uno o pocos eventos de colonización, siendo como tal considerados laboratorios naturales para estudiar la evolución. Tras la colonización inicial, tiene lugar un proceso rápido de diversificación el cual genera altos niveles de endemismo

y diferenciación a nivel eco-morfológico[20–22]. Una de las primeras radiaciones adaptativas estudiadas, y que se sigue explorando hoy en día a nivel genómico, fue descrita por el mismo Charles Darwin en las aves conocidas como pinzones de Darwin, las cuales conforman un grupo de especies que presentan diferencias en el tamaño y forma del pico, resultado de su adaptación a diferentes fuentes de alimento[1,23]. En este proceso, a partir de unos pocos individuos colonizadores se generan distintas especies con una gran variedad de diferencias morfológicas, originadas a pesar de presentar bajos niveles de divergencia genética. No obstante, el papel relativo de la selección natural y de otras fuerzas evolutivas en las radiaciones adaptativas es hoy en día asunto de debate científico[24]. En este contexto, el análisis comparativo de eventos independientes de adaptación tras una colonización inicial, y la subsecuente radiación en archipiélagos, tanto en una misma isla como entre islas, es una aproximación muy prometedora en estudios de genómica evolutiva dado que puede proporcionar nuevos conocimientos sobre los procesos evolutivos que generan diversidad biológica[25,26].

1.4 Convergencia y predictibilidad en la evolución

En el proceso adaptativo a condiciones ambientales concretas, uno de los eventos más llamativos y estudiados son los casos de semejanzas fenotípicas entre distintas especies que no comparten un ancestro común de forma directa, conocido como convergencia fenotípica[27]. De hecho, la evolución convergente es una de las principales evidencias sobre el papel de la selección natural en la diversificación. Uno de los ejemplos de convergencia más estudiados en la literatura es el de los peces marinos conocidos como espinosos (*stickleback*). Estos peces han desarrollado los mismos caracteres fenotípicos en múltiples ocasiones de forma independiente como resultado de su adaptación a ambientes similares. Entre estas adaptaciones se encuentran reducciones en estructuras como la coraza corporal y apéndices pélvicos (involucrados en la adaptación contra depredadores) en poblaciones que invadieron de forma independiente lagos de agua dulce, caracterizados por tener un número reducido de depredadores[28–31].

El conocimiento de la base molecular en estos eventos de convergencia ofrece la oportunidad de estudiar cómo de predecible es el proceso evolutivo. El análisis genómico de escenarios evolutivos paralelos resultados de un origen independiente, permite estudiar cuales han sido las soluciones moleculares explotadas de forma repetida dada su importancia en el proceso adaptativo, y por tanto susceptibles al efecto de la selección natural. A su vez, este tipo de estudios también permite

caracterizar a qué niveles jerárquicos se producen estos cambios de forma repetida, ya sean funciones génicas, vías metabólicas, mismos genes o incluso mismas variantes de aminoácidos o nucleótidos[27].

Las variantes compartidas entre distintas especies con fenotipos convergentes pueden tener distintos orígenes[32]: i) las mutaciones pueden haber ocurrido *de novo* en cada uno de los linajes de forma independiente; ii) las variantes ya estaban presentes segregando en la población ancestral (ancestro común más reciente compartido entre las especies actuales) y han sido fijadas de forma independiente en cada linaje; iii) la mutación ha tenido lugar en un linaje y la presencia de flujo génico entre las especies ha dado lugar a la introgresión adaptativa de esta variante en cada linaje. En particular, estudios recientes sugieren que las variantes genéticas segregando en poblaciones ancestrales habrían tenido un papel fundamental en procesos rápidos de especiación, como las radiaciones adaptativas[31,33–35].

2 El sistema quimiosensorial

2.1 El sistema quimiosensorial y su papel en la adaptación

Unos de los ejemplos más ilustrativos de adaptación es el del sistema quimiosensorial (SQ), al tratarse de un sistema crítico para la supervivencia y reproducción de los organismos dada su implicación en la detección de alimento, huéspedes, predadores, e incluso en el apareamiento[36,37]. Esta habilidad de reconocer y responder de forma distintiva a estímulos externos presenta ventajas adaptativas en los organismos durante el proceso evolutivo. Por tanto, el estudio de genes involucrados en la percepción sensorial ofrece un marco ideal para profundizar nuestros conocimientos en el papel de la selección natural en la adaptación molecular.

La quimiopercepción es posiblemente el sentido más antiguo, presente en todos los organismos tanto unicelulares como multicelulares del planeta, y, en la mayoría de animales, comprende tanto el gusto como el olfato. En términos generales, el sistema gustativo permite la detección de compuestos solubles (hidrofílicos), mientras que el olfativo reconoce compuestos volátiles en el medio terrestre (hidrofóbicos). La discriminación de estos estímulos en fase gaseosa surgiría como una adaptación en el proceso de colonización del medio terrestre, descrito como terrestrialización. De hecho, dado los tiempos de divergencia entre distintos organismos, la terrestrialización se ha producido de forma independiente en distintos taxones animales, incluyendo un mínimo de tres eventos entre los grandes grupos de artrópodos (Figura 3)[38,39]. Sin embargo, algunos artrópodos acuáticos, como los crustáceos, también presentan sistema olfativo sugiriendo así la presencia de estos mecanismos moleculares en el ancestro común de los artrópodos[40]. No obstante, el proceso de terrestrialización implicaría nuevos desafíos y requerimientos para el SQ, viéndose sometido a distintas presiones selectivas en los distintos linajes de artrópodos durante su adaptación de forma independiente al medio terrestre.

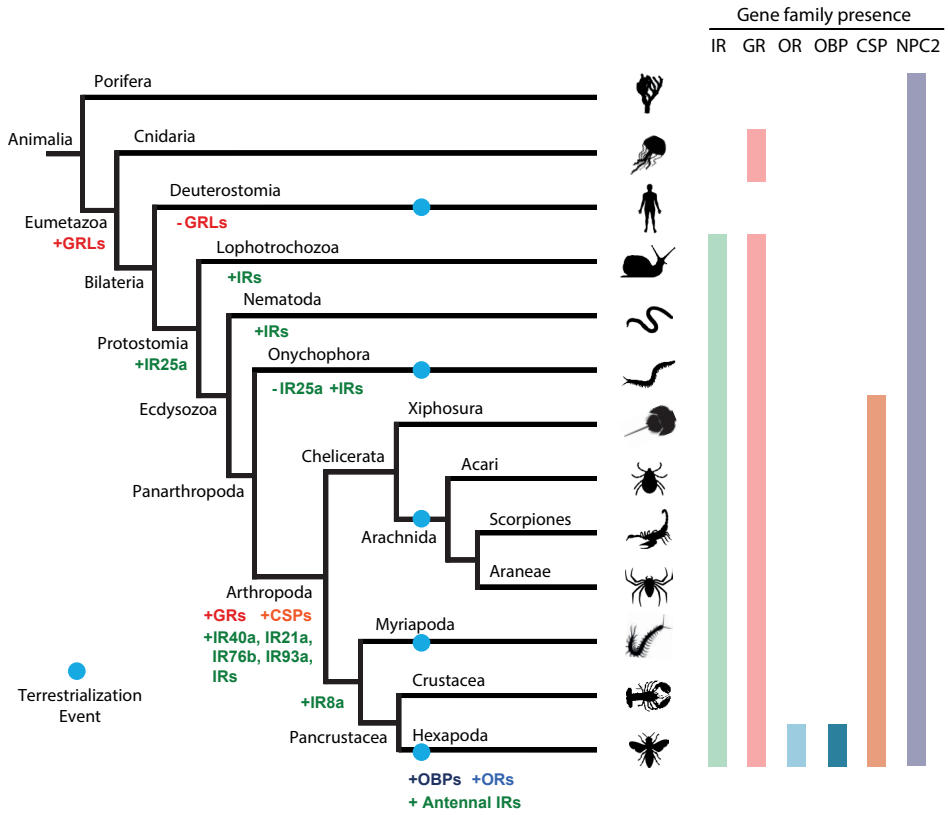


Figura 3. Historia evolutiva de las principales familias del SQ en artrópodos. La presencia o ausencia de las distintas familias multigénicas en cada linaje se indica en columnas (parte de la derecha). Sobre las ramas se representa la aparición (+) o pérdida (-) de una familia o alguno de sus miembros[41–43]. IR: *Ionotropic receptor*; GR: *Gustatory receptor*; OR: *Odorant receptor*; OBP: *Odorant binding protein*; CSP: *Chemosensory protein*; NPC2: *Niemann-Pick C2 protein*. Los puntos azules denotan colonizaciones independientes del medio terrestre en cada linaje. La relación filogenética entre los quelicerados estudiados es la más soportada actualmente, y asume la monofilia de arácnidos[44].

Las proteínas involucradas en la quimiopercepción son codificadas por diversas familias multigénicas que cuentan con un gran número de genes (entre un 1-5% del total de genes en un genoma). Durante el transcurso de la evolución, distintas familias implicadas en el SQ han evolucionado de forma independiente a partir de moléculas diferentes para adquirir la misma función (cooptación), como en insectos y mamíferos. No obstante, la semejanza en la organización y estructura del SQ en ambos grupos de organismos constituye uno de los ejemplos más espectaculares de convergencia evolutiva a nivel molecular[45,46].

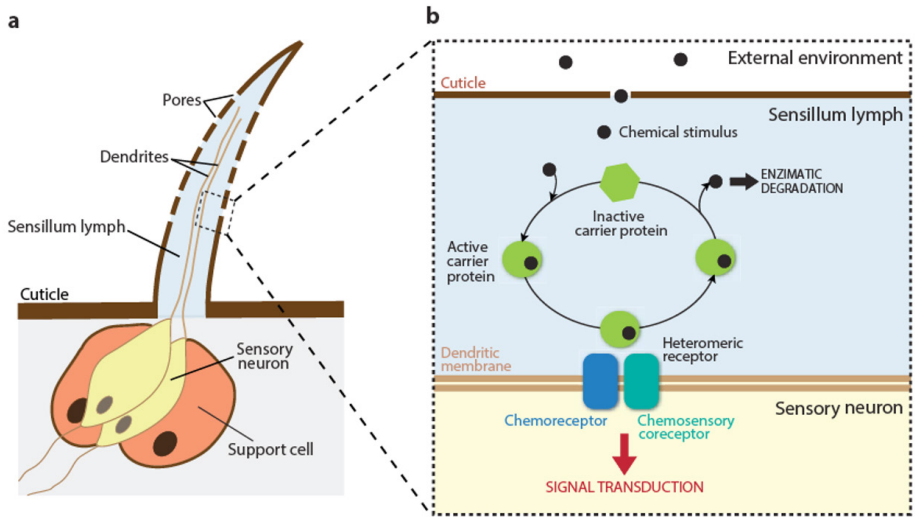


Figura 4. a) Representación esquemática de la estructura general de un pelo olfativo (sensilio) en insectos. Los sensilios gustativos presentan una estructura similar, con la diferencia de que presentan un solo poro en el extremo apical. **b**) Esquema general simplificado del modelo propuesto de los primeros pasos moleculares en la vía de señalización quimiosensorial en insectos. Las proteínas de unión a ligando (*carriers*) se unen al estímulo químico y lo transportan a las proximidades del quimiorreceptor donde se inicia la transducción de señal para el reconocimiento del estímulo olfativo o gustativo. Adaptado de Sánchez-Gracia et al.[47].

2.2 El sistema quimiosensorial periférico en artrópodos

En artrópodos, las primeras etapas en la vía de señalización quimiosensorial tienen lugar en los sensilios, unas estructuras especializadas con forma de pelo. En los sensilios se detecta la señal química y se transmite en forma de señal eléctrica a los glomérulos olfativos donde, posteriormente, se envía al cerebro para su reconocimiento e interpretación (Figura 4). Existe una gran variabilidad en la morfología y organización de estas estructuras quimiosensoriales entre los grandes grupos de artrópodos. En insectos, claramente el grupo mejor estudiado, los sensilios se distribuyen a lo largo de distintos apéndices y difieren según si están involucrados en funciones olfativas o gustativas. Por ejemplo en *Drosophila*, los sensilios olfativos se encuentran en las antenas y palpos maxilares, mientras que los gustativos están presentes en múltiples localizaciones como las patas y alas, y la probóscide[48,49]. En quelicerados, las estructuras quimiosensoriales pueden localizarse principalmente en los pedipalpos o el primer par de patas, aunque también se encuentran en el resto de extremidades y otras partes del cuerpo[50–55]. Sin embargo, dentro de este gran grupo de organismos hay variación en la localización de los glomérulos olfativos en distintas estructuras apendiculares (Figura 5a);

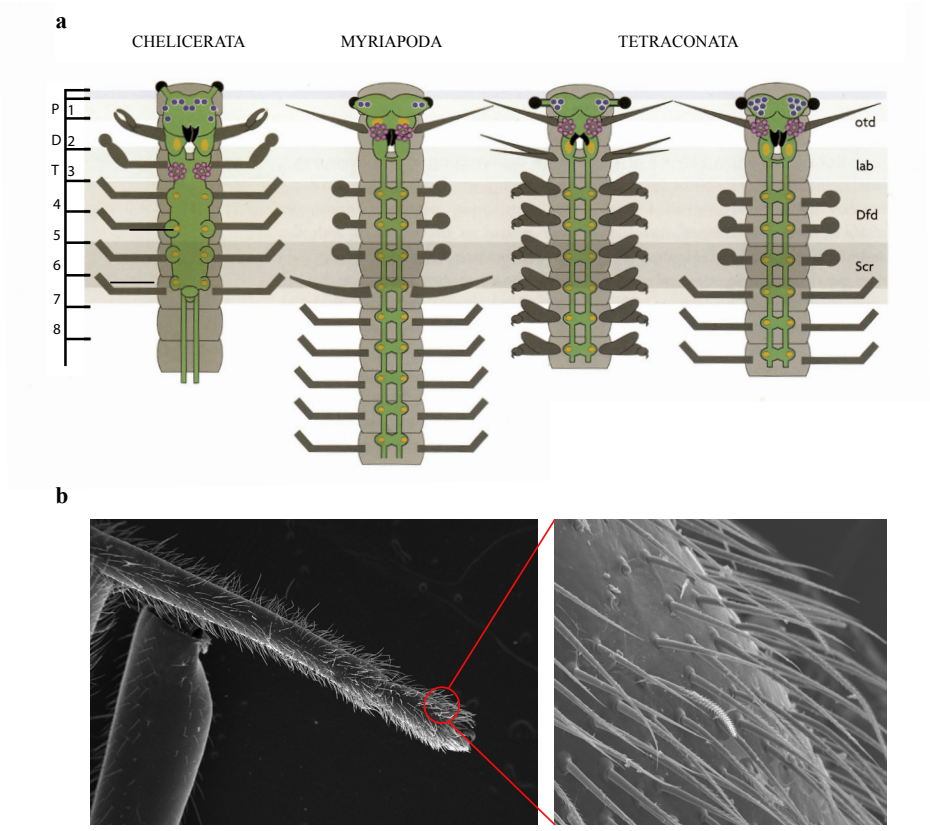


Figura 5. a) Distribución de la expresión génica (*otd*, *lab*, *Dfd* y *Scr*) en distintos segmentos de quelicerados, miriápodos y tetraconados (crustáceos e insectos), evidenciando la homología apendicular en estos grupos de artrópodos. El segundo segmento en mandibulados (miriápodos y tetraconados) está equipado con estructuras olfatorias (que posee las antenas en miriápodos e insectos, y las anténulas en crustáceos) representadas en rosa. El apéndice homólogo en quelicerados son los quelíceros. Sin embargo, los glomérulos olfativos están localizados en el tercer segmento (que incluye los pedipalpos), y/o en el cuarto segmento (primer par de patas), presentando diferencias entre los distintos taxones de quelicerados. Adaptado de Strausfeld[60]. b) Imagen de microscopio electrónico a x50 y x430 de las estructuras quimiosensoriales localizadas en el primer par de patas en la araña *Dysdera silvatica*. Imágenes tomadas por Cristina Frías-López.

en arañas, por ejemplo, se pueden encontrar sensilios de forma predominante en el primer par de patas y pedipalpos (Figura 5b). Finalmente, los crustáceos y los miriápodos perciben los estímulos químicos a través de sensilios localizados en las antenas y en otras zonas del cuerpo, principalmente en patas y apéndices bucales[56–59]. A pesar de estas diferencias, cabe destacar que existe homología entre los apéndices de los distintos grupos de artrópodos sugiriendo que cada grupo ha desarrollado estas estructuras a partir de las ya existentes en su ancestro común (Figura 5a).

2.3 El sistema quimiosensorial en artrópodos: familias multigénicas

Las primeras familias multigénicas implicadas en la quimiopercepción se identificaron en vertebrados y fueron los receptores olfativos que pertenecen a la superfamilia de receptores *G-protein-coupled* (GPCRs), también involucrados en el SQ de nematodos. Sin embargo, en artrópodos, la función de estas proteínas no está relacionada con la detección de estímulos químicos, sino que otras familias han sido cooptadas para esta función[61,62]. Entre los artrópodos, los insectos han sido los más estudiados a nivel molecular, especialmente *Drosophila melanogaster*. No obstante, hasta recientemente, y gracias a la entrada en la era genómica que ha puesto a disposición genomas de distintas especies de artrópodos no insectos, no se ha comenzado a explorar la historia evolutiva de las familias del SQ en el filo de artrópodos al completo. Las familias multigénicas implicadas en el SQ se pueden clasificar en dos grandes tipos, los quimiorreceptores y las proteínas de unión a ligando (Figura 4b).

2.3.1 Quimiorreceptores

Las proteínas descritas como quimiorreceptores se encuentran ancladas en la membrana de las neuronas receptoras localizadas en los sensilios (Figura 4b). En *Drosophila*, las principales superfamilias que modulan las respuestas sensoriales son los receptores olfativos (ORs), gustativos (GRs), y los receptores ionotrópicos (IRs). Los ORs y GRs están involucrados en la respuesta sensorial al olfato y gusto, respectivamente, y pertenecen a una superfamilia de quimiorreceptores (de 400 aminoácidos de longitud aproximadamente) con siete dominios transmembrana (7TM) y con una topología inversa respecto a los receptores GPCRs (Figura 6a). Estos quimiorreceptores de insectos no presentan homología con los GPCRs descritos en vertebrados, y por tanto tienen un origen independiente. Sin embargo, ORs y GRs sí que tienen un origen común, siendo los ORs un linaje más reciente que se originó a partir de GRs, compartiendo ciertas similitudes estructurales y funcionales[63]. Se ha observado que algunos de los ORs no son funcionales por sí mismos, sino que requieren la formación de complejos heteromultiméricos con el co-receptor ORCO (DmelOR83b). De forma similar, tanto GR21a como GR63a son necesarios para la detección CO_2 en *Drosophila*[64,65]. El número de genes de estas familias presenta una gran variación entre insectos, donde se han identificado desde repertorios con 5 copias hasta más de 400 según la especie, con pocos ortólogos

conservados a lo largo del subfilo[66]. Entre los genes más conservados en los insectos se encuentran el co-receptor ORCO, y los GRs involucrados en la percepción de azúcar y CO_2[67–69].

El origen y evolución de estas dos familias es aún asunto de debate en la comunidad científica. Durante los últimos años han surgido distintas hipótesis sobre la aparición del sistema ORCO/OR. Una de las hipótesis propone el origen de este sistema después de la terrestrialización de los insectos, como consecuencia de la adaptación al vuelo en algunos linajes[70]. Sin embargo, tras un estudio más extenso a nivel genómico en taxones ubicados en la base de la clase Insecta y subfilo Hexapoda, se encontraron repertorios de genes ORCO/OR tanto en insectos voladores como no voladores, aunque no en hexápodos no insectos. De esta forma se descartaría la aparición de los ORs como adaptación al vuelo, sugiriendo que su origen es resultado de una innovación en el ancestro de los insectos, probablemente como adaptación a la colonización del medio terrestre (Figura 3)[69].

Paralelamente, se han descrito miembros de la familia de los GRs tanto en todos los subfilos de artrópodos[41,71–74], como en especies de otros filos de metazoos[41,43,75], datando así su origen al inicio de la evolución del reino Animalia (subreino Eumetazoa; Figura 3). Los miembros de esta superfamilia en organismos externos al filo de artrópodos han sido denominados como GR-*like* (GRL) debido a su similitud tanto a nivel de secuencia como estructural con los GRs. No obstante, el número de GRL varía entre 2 y 18 copias en los organismos estudiados[43]; este número reducido podría indicar que su función no está implicada en la quimiopercepción, como se ha visto en el erizo de mar y cnidarios donde los GRL están involucrados en procesos de desarrollo embrionario[75]. De hecho, no se han identificado GRL en distintos genomas de deuterostomados, destacando su ausencia en todos los vertebrados estudiados (Figura 3)[43]. En consecuencia, a pesar de la datación de esta familia como antigua y presente en el origen de los animales, su evolución y funcionalidad en los distintos taxones animales dista de ser comprendida. Por tanto, queda aún por resolver si la función quimiosensorial de los GRs ha sido cooptada por los artrópodos, y si esta coopción se ha producido de forma independiente en los distintos linajes de artrópodos, o ya estaba presente en el ancestro de estos organismos.

La familia de los IRs fue identificada más recientemente[76], y han sido caracterizados tanto en respuestas a estímulos volátiles (olfato) como solubles (gusto)[77]. Estos receptores poseen dos dominios extracelulares de unión a ligando y tres dominios transmembrana (Figura 6a), y están relacionados con los receptores ionótropicos de glutamato (iGluRs). Los iGluRs han sido caracterizados como receptores de

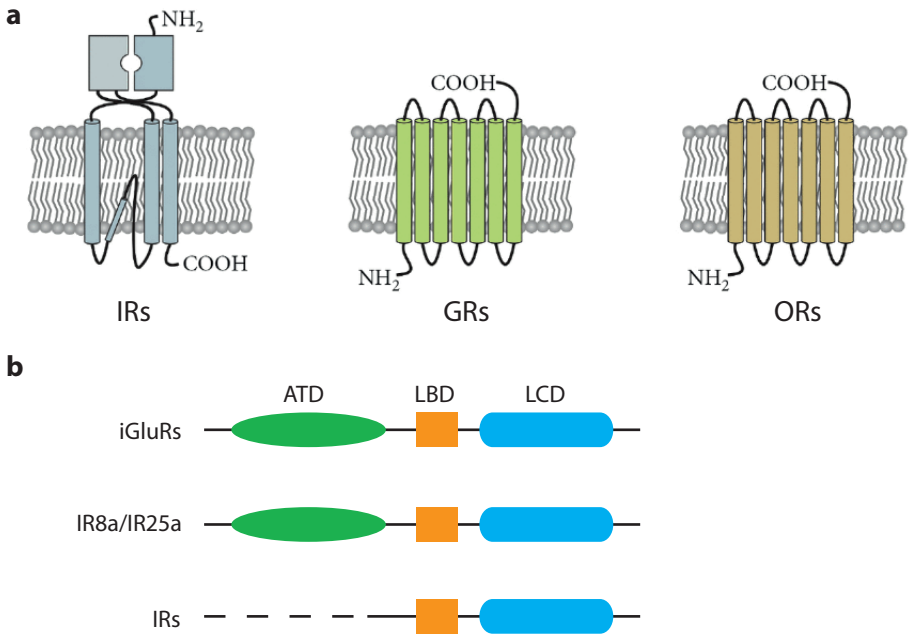


Figura 6. a) Estructura en la membrana de los principales receptores involucrados en la quimiopercepción en artrópodos. Adaptado de Stengl[81]. **b**) Organización de los dominios proteicos en las distintas subfamilias de la familia IR/iGluR. ATD representa el dominio N-terminal siendo localizado únicamente en iGluRs y los correceptores IR8a e IR25a (muy divergente). El dominio de unión a ligando está conformado por dos subdominios (LBD y LCD) presentes en todas las subfamilias. Adaptado de Croset et al.[42].

glutamato, que actúa como neurotransmisor, y su papel es esencial en la transmisión sináptica[78]. Entre los iGluRs, existen tres subfamilias: *α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid* (AMPA), Kainato y *N-methyl-D-aspartate* (NMDA); y se han identificado tanto en vertebrados como invertebrados[79,80]. Los IRs, considerados una nueva subfamilia dentro de los IR/iGluRs, son altamente divergentes en secuencia respecto a los iGluRs y han perdido los residuos implicados en el contacto directo con glutamato[76]. Sin embargo, presentan homología detectada a nivel estructural, conservando los dominios de unión a ligando (LBD y LCD) pero no el dominio extracelular N-terminal (ATD; Figura 6b).

Los miembros de la subfamilia de IRs presentan una alta divergencia entre copias[82], al igual que lo observado en GRs y ORs. De hecho, solo unos pocos receptores se encuentran conservados en insectos, entre los cuales están presentes IR8a e IR25a (denominados como IRs conservados), y los receptores IR93a, IR21a, IR40a e IR76b (Figura 3)[42,82]. Algunos de estos receptores, como los IR8a, IR25a e IR76b han sido descritos como co-receptores de otros IRs, destacando su importancia en el reconocimiento de estímulos químicos por los IRs[83,84]. En *D. melanogaster*, los

IRs se subdividen en dos clases: i) los IRs antenales, que se encuentran expresados de forma predominante en neuronas de sensilios localizados en las antenas y se encargan del reconocimiento de una gran variedad de estímulos olfativos[82,85,86], y ii) los IRs divergentes que están expresados en múltiples tejidos, y la función de algunos de estos receptores ha sido relacionada con el sistema gustativo[87,88].

Los IRs se han identificado en protóstomos (subreino animal que comprende principalmente moluscos, nematodos y artrópodos entre otros filos), pero no en organismos externos a este subreino, como en vertebrados (Figura 3)[42,72–74,89]. En este grupo de organismos se ha identificado el co-recepor IR25a junto con otras secuencias divergentes cuyo repertorio varía enormemente entre especies (desde 3 hasta 85 copias). Dadas estas evidencias, existe la hipótesis de que IR25a se originaría en el ancestro común de los protóstomos a partir de los iGluRs y posteriormente divergiría en secuencia y función adquiriendo la capacidad de actuar como receptor de sustratos distintos al glutamato. A su vez, este receptor se duplicaría dando lugar a los IRs divergentes (Figura 3)[41,42]. Los IRs se han encontrado expresados en tejidos quimiosensoriales de organismos externos al linaje de insectos, como en crustáceos y nematodos, indicando que podrían estar implicados en la quimiopercepción en estas especies[90,91]. En resumen, el origen los IRs es antiguo en la evolución, así como podría ser también su función en la quimiopercepción, habiendo evolucionado y diversificado bajo las presiones selectivas específicas en distintos linajes, como pudo ser la terrestrialización en artrópodos. Es importante destacar que la subdivisión entre IRs antenales y divergentes es específica de insectos, y por tanto la evolución y funcionalidad de IRs más allá de hexápodos, especialmente entre los distintos subfilos de artrópodos, dista aún de ser comprendida.

Además de los receptores codificados por las familias de los GRs, ORs e IRs, se han descrito otras familias multigénicas cuyos miembros también participan en la respuesta quimiosensorial. Tal es el caso de la familia de Deg/ENaCs, también conocidos como PPKs (*pickpocket protein*), donde un reducido número de proteínas han sido identificadas como receptores gustativos en *Drosophila*[92,93]. Además, las proteínas pertenecientes a la familia de las CD36, denominadas como proteínas de membrana de neuronas sensoriales (SNMP), también han sido descritas como co-receptores olfativos y relacionadas con la detección de feromonas[94,95]. Ambas familias, PPKs y CD36/SNMP, están presentes en todos los organismos del reino animal y desarrollan funciones esenciales como el mantenimiento de la salinidad y homeostasis celular (PPKs), o el reconocimiento y transporte de lípidos (CD36)[96,97]. Sin embargo, se desconoce el origen de la innovación evolutiva en miembros de estas familias para la detección de estímulos químicos, dada la ausencia de estudios en artrópodos no insectos de estas familias, tanto a nivel genómico como funcional.

2.3.2 Proteínas de unión a ligando

Además de los quimiorreceptores, la detección de estímulos químicos es mediada, tanto en insectos como en invertebrados, por proteínas globulares solubles expresadas en los órganos del SQ. En términos generales, estas proteínas globulares se encuentran secretadas en el espacio acuoso de los sensilios quimiosensoriales y juegan un papel importante en la unión y transporte de estímulos químicos (se cree que solubilizan las moléculas hidrofóbicas) y la consecuente activación de los quimiorreceptores específicos[49,98]. En vertebrados, estos polipéptidos están codificados por la familia multigénica denominada como *Odorant-Binding Proteins* (*vertebrate* OBPs) que pertenece a la superfamilia de las lipocalinas[99]. En insectos, se han caracterizado principalmente dos familias, las OBPs y las *Chemosensory Proteins* (CSPs)[100]. A pesar de compartir el nombre, las OBPs descritas en vertebrados e insectos no comparten ninguna similitud estructural y no están relacionadas, siendo moléculas no homólogas cooptadas de forma independiente durante la evolución de estos organismos, como en el caso de las GRs/ORs y GPCRs.

Las OBPs y CSPs descritas en insectos codifican proteínas globulares de entre 100 -150 aminoácidos y presentan similitud estructural, habiendo sido considerados como homólogos remotos[100]. Las OBPs presentan un patrón conservado de seis cisteínas, que es crítico para su funcionalidad dado que permite estabilizar su plegamiento globular mediante tres puentes disulfuro[101]. Existen diversas subfamilias de las OBPs en insectos, clasificadas principalmente por diferencias tanto en el perfil de cisteínas como a nivel filogenético: las PBP/GOBP, *minus*-C, *plus*-C, diméricas, ABPI, ABPII, CRLBP y D7[102,103]. En el caso de las CSPs, sus proteínas también exhiben un patrón conservado de cuatro cisteínas, pero presentan un plegamiento distinto al de las OBPs[104,105]. La estructura tridimensional resultante del plegamiento de las proteínas de ambas familias revela una región hidrofóbica de unión a ligando, lo que permitiría el transporte de estímulos volátiles, típicamente hidrofóbicos[106]. No obstante, tanto OBPs como CSPs se expresan también en órganos no quimiosensoriales de insectos e, incluso, algunos miembros de las CSPs participan en otros procesos biológicos como el desarrollo embrionario o la regeneración, por lo que también estarían involucrados en otras funciones distintas a la quimiopercepción, y que quizás fuesen las funciones ancestrales de estas familias[107–109].

Las OBPs han sido descritas únicamente en hexápodos, mientras que las CSPs se han encontrado en artrópodos no insectos, pero no en otros filos (Figura 3)[100,106]. De este modo, Vieira y Rozas[100] formularon la hipótesis de que las OBPs se originaron

a partir de CSPs (ya presentes en el ancestro de artrópodos) durante la evolución del linaje de hexápodos. Sin embargo, dado el bajo número de CSPs encontrados en artrópodos no insectos (1 o 2 copias), y que su función puede no estar involucrada en la quimiopercepción de forma exclusiva, es muy probable que estas proteínas tuviesen algún papel no relacionado con el SQ en el ancestro de los artrópodos y algunos miembros hayan sido reclutados para la detección de estímulos químicos en hexápodos. Por tanto, dadas las evidencias actuales, se desconoce que familias podrían estar implicadas en la solubilización y transporte de estímulos químicos en los linajes de artrópodos, a parte de los insectos.

Con el objetivo de identificar nuevas familias de proteínas globulares implicadas en la quimiopercepción en artrópodos no insectos, Pelosi y colaboradores[106] llevaron a cabo un análisis bioinformático para encontrar posibles candidatos que tuviesen características similares a las OBPs y CSPs de insectos. Los criterios incluidos en su búsqueda de familias multigénicas consistieron en: i) la presencia al menos de 12 genes en una especie para permitir el reconocimiento de distintos estímulos químicos, ii) las proteínas deben ser cortas y solubles, iii) la disposición de una región hidrofóbica de unión a ligando en la estructura tridimensional de la proteína y, iv) la estructura globular de la proteína debe ser estable tanto a altas temperaturas como a agentes químicos y proteólisis. Sorprendentemente, encontraron una familia multigénica en el ácaro *Ixodes scapularis* que cumplía todas las características impuestas, las *Nieman-Pick proteins, type C2* (NPC2).

Las NPC2 se han localizado en todas las especies del reino animal estudiadas hasta la fecha, estando altamente conservadas en vertebrados dónde existe una única copia por especie (Figura 3)[106]. La función de esta proteína en mamíferos está relacionada con la unión y el transporte de colesterol y lípidos[110]. No obstante, el número de copias se encuentra expandido en artrópodos (entre 2 y 14 copias) y su función se desconoce hasta la fecha incluso en insectos. De hecho, se han encontrado varias copias de NPC2 expresadas en las antenas de hormigas y abejas[106,111], soportando la hipótesis de su posible rol en la quimiopercepción. Sin embargo, el papel de estas proteínas en el SQ dista de ser comprendido, requiriendo un estudio en un mayor número de artrópodos no insectos tanto a nivel genómico comparativo como funcional.

Como resumen, los avances durante la última década sobre el origen y evolución de las distintas familias de SQ descritas en insectos han sido significativos, permitiendo conocer los mecanismos moleculares más probables implicados en la quimiopercepción en el ancestro de los artrópodos. No obstante, aún se

desconocen las soluciones moleculares que han cooptado otros grandes grupos de artrópodos, como es el caso de los quelicerados, durante el proceso de adaptación al medio terrestre. Por ejemplo, se desconoce cuáles son las familias implicadas en la percepción de estímulos olfativos que han reclutado los quelicerados como alternativa al sistema OR/OBP descrito en insectos.

2.4 Origen y evolución de las familias multigénicas

El número de miembros de cada familia del SQ difiere ampliamente no sólo entre los grandes grupos de artrópodos, sino también entre especies cercanas[42,100]. La comparación de los repertorios de estas familias multigénicas en distintas especies mediante estudios de genómica comparativa ha mostrado que la dinámica evolutiva que mejor se ajusta al incremento y disminución en el número de copias observado es el modelo denominado como evolución por nacimiento y muerte (Figura 7a). Este modelo explica que los incrementos en el número de genes son debidos, principalmente, a duplicaciones génicas originadas por entrecruzamiento desigual en linajes específicos, mientras que las pérdidas se producen por deleción o pseudogenización[112].

Tras un evento de duplicación génica se generan dos copias idénticas (parálogos) de un mismo gen donde, consecuentemente, puede existir una redundancia funcional que conlleve a una relajación de su constricción funcional en alguna de las copias. Como resultado, los parálogos podrán divergir gradualmente en secuencia por la acumulación independiente de mutaciones, y su destino dependerá de distintos mecanismos evolutivos (Figura 7b). Uno de los posibles destinos evolutivos es la acumulación de mutaciones deletéreas, resultando en la pérdida de función en una de las copias, conocido como pseudogenización. La nueva copia puede adquirir también un cambio nucleotídico que confiera una nueva función, como puede ser el reconocimiento de un nuevo odorante útil para el individuo, proceso denominado como neofuncionalización. En este caso, la nueva copia se encontraría bajo una presión selectiva que ocasionaría su mantenimiento y posterior fijación en la población (selección positiva). Por el contrario, ambos genes pueden divergir y adquirir funciones complementarias, como por ejemplo, la expresión específica en distintos tejidos, siendo así retenidos en el genoma bajo la selección purificadora, debido a que las dos copias se requieren para realizar la función ancestral (subfuncionalización). Finalmente, puede darse el escenario donde ambas copias mantendrían la misma funcionalidad que la copia original si, por ejemplo, se requiere de una mayor síntesis de proteína (Figura 7b)[113].

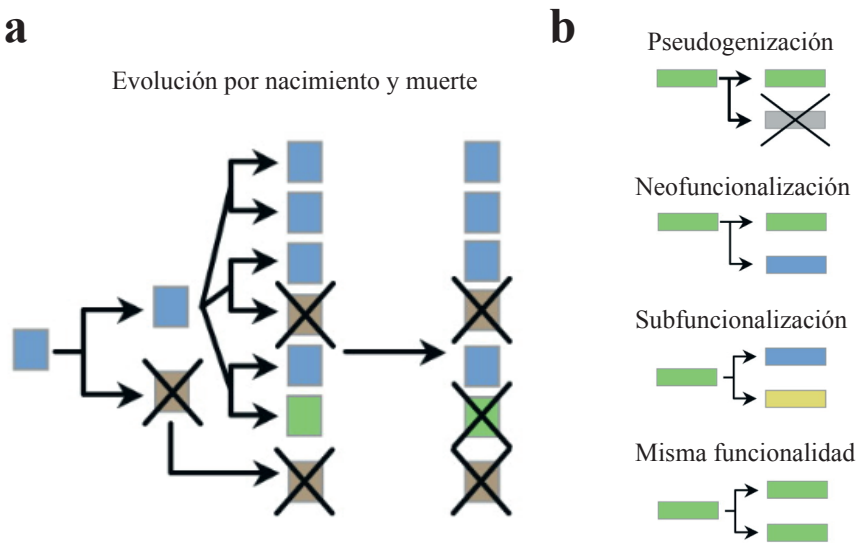


Figura 7. a) Modelo de evolución por nacimiento y muerte de genes: las nuevas copias se generan por duplicación, divergen gradualmente y, eventualmente, se pierden. **b**) Destinos evolutivos de las copias originadas por duplicación: pseudogenización (una de las copias perderá su funcionalidad por acumulación de mutaciones deletéreas); neofuncionalización (una de las copias divergirá adquiriendo una nueva función); subfuncionalización (ambas copiar divergirán adquieriendo funciones complementarias); las copias no presentarán divergencia funcional, manteniendo así la misma función. Los cuadros representan genes y cada color indica una función distinta. Adaptado de Conrad y Antonarakis[114].

Como resultado de la sucesión de eventos de duplicaciones y pérdidas de genes durante la evolución, además de generar una gran variación en el repertorio de las familias multigénicas, un gen en particular podría no tener representante ortólogo en otras especies, o poseer múltiples copias equivalentes incluso en especies cercanas. Esta dinámica limita la capacidad de realizar inferencias funcionales en los miembros de familias multigénicas entre especies, debido a las complejas relaciones de homología producidas por el proceso de ganancia y pérdida de genes. No obstante, esta plasticidad evolutiva tiene un papel importante tanto en la evolución a nivel genómico como en la adaptación, dada la capacidad de la selección natural para que actúe sobre las nuevas copias, lo que puede conllevar a procesos de innovación genética.

3 Organismos de estudio: Quelicerados

Los quelicerados son el segundo subfilo de artrópodos con más especies descritas (cerca de 100,000), siendo únicamente superado por los insectos. Su origen data del Cámbrico, hace unos 530 millones de años (Ma), estimado tanto a partir de inferencias con el registro fósil como a nivel molecular[115,116]. Por lo tanto, la colonización del medio terrestre (380-420 Ma) se produjo de forma independiente en los grandes grupos de artrópodos (quelicerados, hexápodos, crustáceos y miriápodos; Figura 3)[115–117]. El subfilo de quelicerados presenta una gran diversidad de animales terrestres, incluyendo arañas, ácaros, escorpiones y opiliones entre otros; además de organismos acuáticos como los xifosuros (cangrejo de herradura) y los picnogónidos (arañas de mar). Las especies pertenecientes a quelicerados, además de su gran capacidad de adaptación y su enorme biodiversidad, exhiben características que les confiere una gran importancia económica y médica. Tal es el caso de las posibles aplicaciones biotecnológicas que ofrece el estudio de la seda en arañas, o los potentes venenos descritos en arañas y escorpiones[118–121]. A su vez, muchas de estas especies interaccionan directamente con el ser humano, siendo algunos organismos, especialmente entre los ácaros, vectores de enfermedades o plagas de cultivos[74,122].

Actualmente existe una gran incertidumbre sobre las relaciones filogenéticas entre los distintos linajes de quelicerados, especialmente con respecto al soporte de la monofilia de arácnidos. Ello es debido, principalmente, a la falta de resolución en la posición de los ácaros y xifosuros, y que podría implicar relaciones parafiléticas en el grupo de los arácnidos[123,124]. La clase Arachnida comprende distintos grupos, todos ellos con un estilo de vida exclusivamente terrestre. En consecuencia, la relación parafilética entre arácnidos implicaría que la colonización del medio terrestre hubiera ocurrido múltiples veces de forma independiente en este grupo de organismos[123]. No obstante, otros datos moleculares recientes sí que soportan la monofilia de arácnidos sugiriendo que hubo un único evento de terrestrialización dentro de quelicerados (Figura 3)[44], aunque el debate aún sigue abierto.

3.1 El género *Dysdera*

Las arañas pertenecen al orden Araneae y comprenden un grupo con una gran biodiversidad dentro de artrópodos. Se han descrito aproximadamente unas 45,000 especies, siendo el orden más numeroso en la clase Arachnida[125]. Estas especies son depredadores dominantes en la mayoría de ecosistemas terrestres, presentando un gran número de estrategias para capturar a sus presas. Dentro de este orden, las arañas del género *Dysdera* Latreille 1804 (Araneae, Dysderidae) en las Islas Canarias presentan una de las radiaciones más espectaculares en arácnidos. Se han descrito aproximadamente unas 250 especies de este género distribuidas en la región mediterránea[125,126]. También se encuentran en islas de Macaronesia (comprende cinco archipiélagos de origen volcánico localizados en el atlántico norte cercanos al continente africano, entre ellos las Islas Canarias), representando la región más al oeste en su distribución geográfica. Sorprendentemente, 47 especies han sido catalogadas como endémicas de las Islas Canarias, comprendiendo así un 20% de la diversidad conocida de este género en una región que representa un 0,1% de su distribución geográfica[126]. Por el contrario, únicamente se conoce una especie endémica en Azores, Islas Salvajes y Cabo Verde, y cinco en Madeira. Dentro de las Islas Canarias, se han descrito entre dos y tres eventos de colonización independientes, observando clados distintos entre las especies de las islas localizadas en el este y oeste[127,128]. Esta colonización de las Islas Canarias por el género *Dysdera* se ha estimado que tuvo lugar poco después de surgir las primeras islas (alrededor de 20 Ma; M.A. Arnedo comunicación personal).

Las arañas del género *Dysdera* son cazadoras nocturnas y durante el día se encuentran cubiertas en capullos de seda bajo piedras, cortezas y hojarasca, llegando algunas especies incluso a habitar en cuevas[128]. Este género destaca entre las arañas dado que algunas especies han desarrollado una especialización trófica (denominado como estenofagia), mientras que otras son completamente generalistas. Esta especialización consiste en la alimentación, de forma facultativa o incluso obligatoria en algunas *Dysdera*, de isópodos terrestres (Crustacea: Isopoda), una presa que es rechazada por la mayoría de depredadores generalistas[129–131]. Los isópodos terrestres son una presa evitada por los artrópodos debido a sus mecanismos de defensas a nivel morfológico, químico y comportamiento[132,133]. Entre estas defensas destaca su capacidad de enrollarse sobre sí mismas, formando una bola cuando se sienten amenazadas, facilitado por su duro exosqueleto que actúa como armadura protegiéndolos de sus depredadores. A su vez, estos organismos también presentan secreciones externas que generan olores repulsivos e incluso producen indigestión en sus depredadores, además de tener hábitos

nocturnos para evitar ser detectados[133,134]. Por último pero no menos importante, los isópodos terrestres son capaces de acumular altas concentraciones de metales pesados como consecuencia de su adaptación a ecosistemas que presentan elevados niveles de contaminación, siendo utilizados como bioindicadores de suelos contaminados[135,136]. Los metales pesados son fundamentales para las funciones fisiológicas y bioquímicas, pero su ingestión en altas concentraciones es perjudicial para la mayoría de organismos[137]. Esta característica confiere a los isópodos de una gran toxicidad para sus depredadores[138]. De este modo, los isópodos no suelen ser cazados por depredadores generalistas y, entre artrópodos, solo se conocen algunas arañas y hormigas que se han especializado en alimentarse de esta presa[129,139].

La estenofagia en *Dysdera* se ha originado de forma independiente en varias ocasiones, tanto en el continente como en las islas (M.A. Arnedo comunicación personal). Las arañas especialistas de isópodos presentan distintas morfologías en los quelíceros que se asocian tanto con estrategias de captura para superar las defensas de los isópodos, como con preferencias de presa[131]. Además de estas adaptaciones morfológicas y de comportamiento, también se ha observado que las arañas especialistas presentan un mayor crecimiento y asimilación de nutrientes cuando se alimentan de isópodos, en vez de otras presas cazadas por especies generalistas, sugiriendo así una notable adaptación nutricional[140,141]. De este modo, la evolución repetida de la estenofagia, caracterizada principalmente por las morfologías en los quelíceros, sugiere que la segregación de especies según el tipo de presa ha tenido un impacto importante en la gran diversificación del género *Dysdera* en las Islas Canarias[128]. Sin embargo, a pesar de las evidencias morfológicas y experimentales descritas, se desconoce totalmente la base genética y el impacto de las distintas fuerzas evolutivas en esta destacada adaptación.

Previo al inicio de esta tesis doctoral, solo se encontraban disponibles las secuencias genómicas de un reducido número de artrópodos no insectos, con sólo un quelicerado publicado (además de contar con los datos del ácaro *Ixodes scapularis* donde participó nuestro grupo de investigación)[74,122]. A su vez, el número de estudios genómicos de las familias del SQ incluyendo especies de todos los subfilos de artrópodos era muy limitado[42,71,72,100,106]. Durante el desarrollo de este trabajo, se han generado datos transcriptómicos de las arañas del género *Dysdera* y, además, otros grupos de investigación han secuenciado y dispuesto de forma pública para la comunidad científica un gran número de nuevos genomas incluyendo diversas especies de quelicerados. Así, durante esta tesis se han utilizado estos datos transcriptómicos y genómicos para estudiar el origen y evolución de las familias del SQ en este gran grupo de organismos.

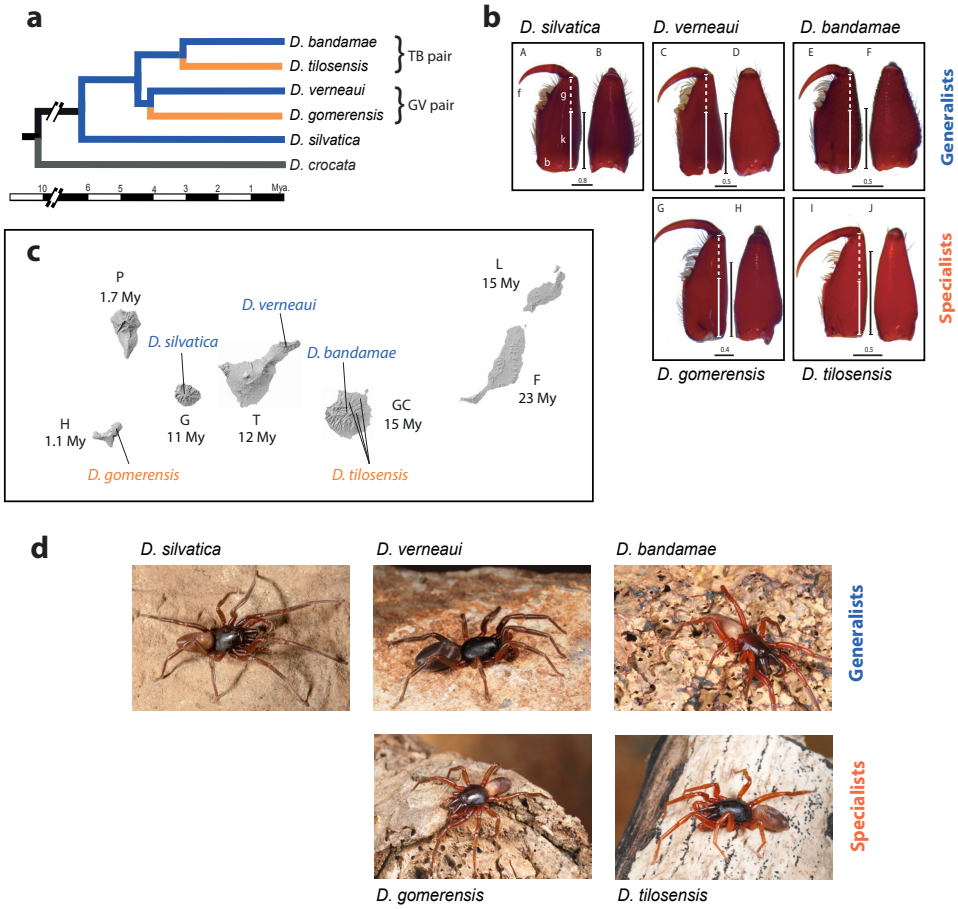


Figura 8. a) Relaciones filogenéticas y tiempos de divergencia (escala en millones de años) entre las especies del género *Dysdera* estudiadas: *D. gomerensis* Strand, 1911 (El Hierro), *D. verneaui* Simon, 1883 (Tenerife), *D. tilosensis* Wunderlich, 1992 (Gran Canaria), *D. bandamae* Schmidt, 1973 (Gran Canaria), y *D. silvatica* Schmidt, 1981 (La Gomera). En color azul y naranja se representan las especies cuya dieta es generalista o especialista (alimentación preferente de isópodos), respectivamente. **b**) Vista ventral y lateral del quelícero izquierdo diseccionado en cada una de las especies (escala en milímetros). Las barras indican la longitud relativa de las distintas partes del quelícero, para destacar las diferencias entre quelíceros estándar (asociados a especies generalistas) y alargados (característicos de especies especialistas). **c**) Mapa de las Islas Canarias incluyendo la localización geográfica donde se capturaron los especímenes de cada especie incluidos en el estudio. La edad aproximada de formación de cada isla (en millones de años) se indica en negro[142]: F: Fuerteventura, L: Lanzarote, GC: Gran Canaria, T: Tenerife, G: La Gomera, P: La Palma y H: El Hierro. **d**) Imágenes de cada especie de *Dysdera* en su hábitat, tomadas por Pedro Oromí.

A su vez, con objeto de comprender la base genética de la adaptación específica observada en las arañas del género *Dysdera*, y proporcionar nuevos datos y conocimiento sobre el continuo debate de cómo de predecible es la evolución

molecular, se diseñó un caso de estudio que incluía dos parejas de especies especialistas-generalistas, con un *outgroup* generalista, endémicas de las Islas Canarias (Figura 8). Estudios previos indican que, con casi toda seguridad, el cambio de dieta se produjo de forma independiente en ambas parejas de especies, a partir de un ancestro generalista, y en localizaciones geográficas separadas. A partir de los datos transcriptómicos, se compararon tanto los perfiles de expresión génica como los patrones de constricción selectiva entre especies especialistas y generalistas, con el objetivo de identificar las regiones genómicas involucradas en la adaptación trófica descrita, así como los mecanismos evolutivos implicados en este proceso.

Objetivos

La adaptación al medio juega un papel vital en nuestro conocimiento y comprensión del proceso de especiación y de la biodiversidad. El objetivo fundamental de esta tesis es el de profundizar en el conocimiento de los procesos adaptativos, y en particular, el papel de la selección natural en la adaptación a nivel molecular. Para ello, se han estudiado las familias multigénicas del sistema quimiosensorial con el fin de aportar conocimiento sobre su origen y evolución y su posible participación en la adaptación al medio terrestre de los distintos subfilos de artrópodos. De forma paralela, y a una escala temporal inferior, se ha estudiado la base genética del proceso de especialización trófica en las arañas del género *Dysdera*, con la finalidad de caracterizar y determinar si las soluciones moleculares utilizadas por la evolución se han producido de forma repetida y comprender las radiaciones adaptativas desde un punto de vista genómico.

Los objetivos específicos de esta tesis doctoral han sido:

- Desarrollar metodologías bioinformáticas para la identificación y anotación de familias multigénicas a nivel genómico y transcriptómico.

- Estudiar el origen y evolución de las familias multigénicas del sistema quimiosensorial (FMSQ) en artrópodos:
  - Caracterizar las principales FMSQ en quelicerados mediante el estudio por transcriptómica de los tejidos quimiosensoriales en la araña *Dysdera silvatica*.
  - Estudiar la evolución de las FMSQ mediante el análisis genómico comparativo de 11 genomas de quelicerados.

- Determinar la base genómica de la especialización trófica observada en la radiación adaptativa de *Dysdera* en las Islas Canarias.

Informe de los directores

UNIVERSITAT DE BARCELONA

Dr. Julio Rozas
Catedràtic de Genètica

Departament de Genètica, Microbiologia i Estadística
Facultat de Biologia

Diagonal 643
Edifici Prevosti
08028 Barcelona
Spain

Tel. +34 93 4021495
Fax. +34 93 4034420
jrozas@ub.edu
www.ub.edu/molevol/julio

Informe signat del director de tesi del factor d'impacte dels articles publicats. En cas que es presenti algun treball en coautoria, caldrà incloure també un informe del director de la tesi signat, en què s'especifiqui exhaustivament quina ha estat la participació del doctorant/a en cada article, i si algun dels coautors d'algun dels treballs presentats en la tesi doctoral ha utilitzat, implícitament o explícitament aquests treballs per a la l'elaboració de la tesi doctoral

---

El Drs. **Julio Rozas i Alejandro Sánchez-Gracia,** directors de la Tesi Doctoral elaborada pel Sr. Joel Vizueta Moraga, amb el títol "**Genómica de la adaptación en artrópodos: estudio del sistema quimiosensorial y de la radiación del género *Dysdera* (Araneae) en Canarias**"

INFORMEN

Que la tesi doctoral està elaborada com a compendi de 5 publicacions amb dades originals (publicacions 1-4 en el cos central de la tesi), i la 5èna a l'apèndix:

Publicacions:

1. Vizueta, J., Sánchez-Gracia, A. and Rozas, J. 2019. BITACORA: A comprehensive tool for the identification and annotation of gene families in genome assemblies. *bioRxiv* https://doi.org/10.1101/593889
   Preparat per enviar a enviar a una revista amb *peer review*.

2. Vizueta, J., Frías-López, C., Macías-Hernández, N., Arnedo, M. A., Sánchez-Gracia, A. and Rozas, J. 2017. Evolution of chemosensory gene families in arthropods: Insight from the first inclusive comparative transcriptome analysis across spider appendages. *Genome Biol. Evol.* **9:** 178-196.
   Factor d'impacte (5 Year Impact Factor): **4.171**. **Q1** dins la categoria de Genetics & Heredity.

3. Vizueta, J., Rozas, J. and Sánchez-Gracia, A. 2018. Comparative genomics reveals thousands of novel chemosensory genes and massive changes in chemoreceptor repertories across chelicerates. *Genome Biol. Evol.* **10:** 1221-1236.
   Factor d'impacte (5 Year Impact Factor): **4.019**. **Q1** dins la categoria de Genetics & Heredity.

4. Vizueta, J., Macías-Hernández, N., Arnedo, M. A., Rozas, J. and Sánchez-Gracia, A. 2019. Chance and predictability in evolution: the genomic basis of convergent dietary specializations in an adaptive radiation. *Mol. Ecol.* **28**: 4028-4045.
   Factor d'impacte (5 Year Impact Factor): **6.614**. **Q1** i **D1** (dades del 2018) dins les categories de Ecology; Evol. Biology.

UNIVERSITAT DE BARCELONA

Dr. Julio Rozas
Catedràtic de Genètica

Departament de Genètica, Microbiologia i Estadística
Facultat de Biologia

Diagonal 643
Edifici Prevosti
08028 Barcelona
Spain

Tel. +34 93 4021495
Fax. +34 93 4034420
jrozas@ub.edu
www.ub.edu/molevol/julio

5. Frías-López, C., Almeida, F. C., Guirao-Rico, S., Vizueta, J., Sánchez-Gracia, A., Arnedo, M. A. and Rozas, J. 2015. Comparative analysis of tissue-specific transcriptomes in the funnel-web spider *Macrothele calpeiana* (Araneae, Hexathelidae). *PeerJ* **3:** e1064.
   Factor d'impacte (5 Year Impact Factor): **2.183**. **Q1** dins la categoría de Multidisciplinary Sciences.

A les publicacions 1-4, el doctorant va realitzar la feina computacional, analítiques i redacció del primer esborrany del manuscrit. La publicació 5 (apèndix), és el resultat d'una col·laboració científica on el doctorant, fent servir eines computacionals o analítiques desenvolupades en la seva tesi doctoral, ha realitzat una part dels anàlisis bioinformàtics. En cap cas s'ha utilitzat, implícitament o explícitament, els treballs presentats en el cos central d'aquesta tesi per a l'elaboració d'una altra tesi doctoral.

Dr. Julio Rozas Liras
Catedràtic de Genètica
Universitat de Barcelona

Alejandro Sánchez-Gracia
Professor Associat de Genètica
Universitat de Barcelona

Capítulos

1

BITACORA: A comprehensive tool for the identification and annotation of gene families in genome assemblies

Actualmente, el proceso de anotación de nuevos genomas es un cuello de botella en genómica, especialmente en el estudio de familias multigénicas en organismos no modelo. A pesar de los avances que se han dado en el desarrollo de metodologías automáticas para realizar la anotación estructural, estas herramientas generan frecuentemente anotaciones erróneas (como genes dobles, quiméricos o parciales) o incluso no son capaces de predecir modelos génicos para diversas copias de familias multigénicas, lo que conlleva a realizar un esfuerzo manual para su correcta anotación. En este artículo se presenta BITACORA, una herramienta bioinformática que integra algoritmos de búsqueda por similitud de secuencia con scripts de Perl para facilitar la subsanación de errores en la anotación, y la identificación de miembros de familias multigénicas en secuencias genómicas que no contaban con anotación estructural. Hemos evaluado BITACORA usando datos de la anotación de dos familias multigénicas involucradas en el sistema quimiosensorial en siete genomas de quelicerados. A pesar de la alta fragmentación relativa de algunos de estos genomas, BITACORA mejoró la anotación en una gran parte de genes de las dos familias estudiadas, y detectó miles de nuevos quimiorreceptores previamente no anotados. Esta herramienta genera ficheros GFF que incluyen los modelos génicos tanto de los genes previamente anotados como las nuevas copias identificadas, y los ficheros FASTA con las proteínas y CDS codificados por estos genes. Estos ficheros pueden ser fácilmente integrados en editores de anotación genómica, facilitando los procesos de anotación semi-automáticos y posteriores análisis evolutivos.

Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

New Results

BITACORA: A comprehensive tool for the identification and annotation of gene families in genome assemblies

Joel Vizueta, Alejandro Sánchez-Gracia, Julio Rozas

doi: https://doi.org/10.1101/593889

Abstract

Gene annotation is a critical bottleneck in genomic research, especially for the comprehensive study of very large gene families in the genomes of non-model organisms. Despite the recent progress in automatic methods, the tools developed for this task often produce inaccurate annotations, such as fused, chimeric, partial or even completely absent gene models for many family copies, which require considerable extra efforts to be amended. Here we present BITACORA, a bioinformatics solution that integrates sequence similarity search tools and Perl scripts to facilitate both the curation of these inaccurate annotations and the identification of previously undetected gene family copies directly from DNA sequences. We tested the performance of the BITACORA pipeline in annotating the members of two chemosensory gene families of different sizes in seven available chelicerate genome drafts. Despite the relatively high fragmentation of some of these drafts, BITACORA was able to improve the annotation of many members of these families and detected thousands of new chemoreceptors encoded in genome sequences. The program generates an output file in the general feature format (GFF) files, with both curated and novel gene models, and a FASTA file with the predicted proteins. These outputs can be easily integrated in genomic annotation editors, greatly facilitating subsequent manual annotation and downstream evolutionary analyses.

Footnotes

- http://www.ub.edu/softevol/bitacora/

BITACORA: A comprehensive tool for the identification and annotation of gene families in genome assemblies

Joel Vizueta*, Alejandro Sánchez-Gracia* and Julio Rozas*

Departament de Genètica, Microbiologia i Estadística and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain

*To whom correspondence should be addressed.

Corresponding authors: jvizueta@ub.edu, elsanchez@ub.edu and jrozas@ub.edu

Running head

BITACORA: A tool for gene family annotation

Introduction

The falling cost of high-throughput sequencing (HTS) technologies made them accessible to small labs, promoting a large number of genome-sequencing projects even in non-model organisms. Nevertheless, genome assembly and annotation, especially in eukaryotic genomes, still represent major limitations (Dominguez Del Angel et al., 2018). The unique genomic characteristics of many non-model organisms, often lacking pre-existing gene models (Yandell & Ence, 2012), and the absence of closely related species with well-annotated genomes, converts the annotation process in a big challenge. The state-of-the-art pipelines for *de novo* genome annotation, like BRAKER1 or MAKER2, allow integrating multiple evidences, such as RNA-seq, EST data or gene models from other annotated species (using for example GeneMark, Exonerate, or GenomeThreader) with *ab initio* gene predictions (from Augustus or SNAP) in order to produce structural annotations of genome sequences (Gremme, Brendel, Sparks, & Kurtz, 2005; Hoff et al., 2016; Holt & Yandell, 2011; Korf, 2004; Lomsadze, Burns, & Borodovsky, 2014; Slater & Birney, 2005; M. Stanke & Waack, 2003; Mario Stanke, Diekhans, Baertsch, & Haussler, 2008). Some of these pipelines, such as BRAKER1, will only report those gene models with evidences. However, the gene models predicted by these automatic tools are often inaccurate, especially those belonging to gene families. Their curation frequently requires the use of additional programs, such as Augustus-PPX (Keller, Kollmar, Stanke, & Waack, 2011), or semi-automatic approaches evaluating the quality of supporting data. This latter task is usually performed in genomic annotation editors, such as Apollo, which give researchers the option to work simultaneously in the same annotation project (Lee et al., 2013).

There are a number of issues affecting the quality of gene family annotations, especially for either old or fast evolving families (Yohe et al., 2019). First, new duplicates within a family usually originate by unequal crossing-over and are found in tandem arrays in the genome, being the more recent duplicates also the physically closest (Clifton et al., 2017; Vieira, Sánchez-Gracia, & Rozas, 2007). This configuration often causes local miss-assemblies that result in the incorrect or failed identification of tandem duplicated copies (i.e., it produces artifact, incomplete, or chimeric genes along a genomic region). Secondly, the identification and characterization of gene copies in medium- to large-sized families tends to be laborious, requiring data from multiple sources, including well-annotated remote homologs and hidden Markov model (HMM) profiles. Certainly, the fine and robust identification and annotation of the complete repertory of a gene family in a typical genome draft is a challenging task that requires important additional efforts, which are very tedious to perform manually.

In order to facilitate this curation task, we have developed BITACORA, a bioinformatics pipeline to assist the comprehensive annotation of gene families in genome assemblies. BITACORA requires of a structurally annotated genome (GFF and FASTA format), and a curated database with well-annotated members of the focal gene families. The program will perform comprehensive BLAST and HMMER searches (Altschul, 1997; Eddy, 2011) to identify putative candidate gene regions (already annotated, or not), combine evidences from all searches and generate new gene models. The outcome of the pipeline consists of a new structural annotation (GFF) file along with their encoded sequences. These output sequences can be directly used to conduct downstream functional or evolutionary analyses, to be included as evidences in other annotation pipelines (BRAKER1 or MAKER2; Hoff et al., 2016; Holt & Yandell, 2011), to improve existing gene models predictions, or to facilitate a fine re-annotation in genome browsers such as Apollo (Lee et al., 2013).

Methods and implementation

Input data files

BITACORA requires: i) a data file with the genome sequences (in FASTA format), ii) the associated GFF file with annotated features (either in GFF3 or GTF formats; features must include both transcript or mRNA and CDS), iii) a data file with the predicted proteins included in the GFF (in FASTA format), and iv) a database (here referred as FPDB database) with the protein sequences of well annotated members of the gene family of interest (focal family; in FASTA format) along with its HMM profile (see Supplementary Material for a detailed description of FPDB construction). Since sequence similarity-based searches are very sensitive to the quality of the proteins in FPDB, it is important to include in this database highly curated proteins from closely related species. This is especially important for the annotation of very old or fast-evolving gene families. Also, the use of a HMM profile increases the likelihood of identifying sequences encoding new members; these profiles can be obtained from external databases (such as PFAM) or build using high quality protein alignments with the program *hmmbuild* (Finn *et al.*, 2014). Before starting the analysis, BITACORA checks whether input data files are correctly formatted; otherwise, it will suggest some format converters distributed with the program (see Troubleshooting section in Supplementary Material).

Curating existing annotations

The BITACORA workflow is divided in three main steps (Fig. 1). The first step consists in the identification of all putative homologs of the FPDB sequences from the focal gene family that are already present in the input GFF file, and the curation of their gene models (referred hereinafter as b-curated (bitacora-curated) gene models or proteins). Specifically, the pipeline launches BLASTP and HMMER searches (Altschul, 1997; Eddy, 2011) against the proteins predicted from the features in the input GFF using the FPDB protein sequences and HMM profiles as queries; the resulted alignments are filtered for quality (i.e. BLASTP hits covering at least two-thirds of the length of query sequences or including at least the 80% of the complete protein used as a subject are retained). The results from both searches are combined into a single integrated result for every single protein (gene model). Then, BITACORA trims the original models based in these combined results, and reports new gene coordinates (b-curated models) in a new updated GFF (uGFF), fixing for example all chimeric annotations. Besides, the proteins encoded by these b-curated models are incorporated to the FPDB (updated FPDB or uFPDB), to be used in an additional search round.

Identifying genomic regions encoding new family members

In the second step, BITACORA uses TBLASTN to search the genome sequences for regions encoding homologs of the proteins included in the uFPDB but not annotated in the uGFF. Overlapping TBLASTN hits, which we would expect to represent a unique exon sequence, are merged into one single alignment. Then, all alignments located in the same scaffold and separated less than the maximum allowed intron distance (indicated by the "intron distance parameter") are connected to obtain a putative single protein coding region (referred hereinafter as b-novel gene models). This step is intended to join coding exons of the same gene based on expected intron distance in the surveyed genome. We provide some scripts to estimate the "intron distance parameter" from the input GFF (see Supplementary Material). Last, to avoid reporting inaccurate b-novel gene models and to identify putative gene fusions among them, BITACORA checks the encoded proteins for the presence of the gene family-specific domain (using the HMM profile in FPDB), and only models having this domain are reported in the final dataset of annotated proteins, tagging those cases that could be the result of a fusion of multiple genes with the label 'Ndom' (being N =>2, denoting the presence of more than one protein family domain in the sequence; see Supplementary Material for more details).

Optional search round and final output

Finally, BITACORA can also be used to perform a second search round using as the input data all proteins obtained in steps 1 and 2 (sFPDB database). This additional step is especially useful for searching remote homologs undetected in the previous steps. The final BITACORA outcome will include therefore, 1) an updated GFF file with both b-curated and b-novel gene models, 2) all non-redundant proteins predicted from these feature annotations (in a FASTA file), 3) two BED files, one with all gene coordinates in the genome sequence and the other with only those regions that encode the novel members of the focal family identified by the program and, 4) all protein sequences found in all steps.

Additional features

BITACORA could be also used in the absence of either a reference genome for the target species (e.g. for transcriptomic studies) or a precompiled GFF (e.g. for non-annotated genomes); in these cases, the input should be a FASTA file with the set of predicted proteins or the genome sequences, respectively (see Supplementary Material for alternative usage modes). With BITACORA, we also distribute a series of scripts to perform some useful tasks, such as estimating intron length statistics from a GFF, converting GFF to GTF format, and retrieving all protein sequences encoded by the features of a GFF file. Furthermore, to better adjust to the particularities of each genome, BITACORA allows the user to specify the values of most important parameters, such as the *E*-value for BLAST and HMMER searches, the number of threads in BLAST runs, or the maximum intron length required to connect putative exons of the same gene.

BITACORA application example

As a demonstration of the performance of BITACORA in a group of genomes of different quality and assembly contiguity, we present the extended results of the annotation of two arthropod chemosensory gene families, the insect gustatory receptor (GR) and the Niemann-Pick type C2 (NPC2) gene families (Pelosi et al., 2014; Robertson, 2015), in a subset of seven chelicerate genomes from those analyzed in Vizueta et al., (2018). For the analysis, we retrieved the data (genome sequences, annotations and predicted peptides) of the scorpions *Centruroides sculpturatus* (bark scorpion, genome assembly version v1.0, annotation version

v0.5.3; Human Genome Sequencing Center (HGSC)) and *Mesobuthus martensii* (v1.0, Scientific Data Sharing Platform Bioinformation (SDSPB)) (Cao et al., 2013); and of the spiders *Acanthoscurria geniculata* (tarantula, v1, NCBI Assembly, BGI) (Sanggaard et al., 2014), *Stegodyphus mimosarum* (African social velvet spider, v1, NCBI Assembly, BGI) (Sanggaard et al., 2014), *Latrodectus hesperus* (western black widow, v1.0, HGSC), *Parasteatoda tepidariorum* (common house spider, v1.0 Augustus 3, SpiderWeb and HGSC) (Schwager et al., 2017) and *Loxosceles reclusa* (brown recluse, v1.0, HGSC). The GR and NPC2 families show very different protein and genomic features. The GR gene family encodes seven-transmembrane receptors of ~400 amino acids long with an average of 2.3 exons per gene in the genome of the spider *Parasteatoda tepidariorum*; the NPC2 proteins are ~150 amino acids long and have an average of 2.6 exons per gene in the same species.

Strikingly, BITACORA uncovered the identification of thousands of new gene models previously undetected in these chelicerate genomes. For instance, BITACORA was able to identify and annotate 1,234 GR encoding sequences in the bark scorpion *Centruroides sculpturatus*, where only 24 proteins were initially identified by the automatic annotation pipelines (Table 1). Globally, BITACORA identified, annotated and curated 3,371 sequences encoding GR proteins in the seven genomes (3,265 of them absent in structural annotations included in the GFF of these genomes). It is largely known that this gene family evolves rapidly in arthropods, both in terms of sequence change and repertory size, encoding in the same genome very recent and distantly related receptors as well as pseudogenes. Since some of these receptors show a very restricted gene expression pattern (expressed in specialized cells and tissues involved in chemoreception), their transcripts are often missing in RNA-seq data sets, which are one of evidences used for the automatic annotation of the genomes (Joseph & Carlson, 2015; Robertson, 2015; Vizueta et al., 2017; Zhang, Zheng, Li, & Fan, 2014). This fact, added to the huge divergence accumulated between many copies (a mixture of age and rapid evolution), probably prevented the automatic annotation of the GRs uncovered by BITACORA.

The members of the NPC2 family, on the contrary, are much more conserved at the sequence level and show higher levels of gene expression in arthropods (Pelosi et al., 2014). As expected, the number of newly identified copies of this family in the seven chelicerate genomes is much lower than in the case of GRs. Even that, BITACORA was able to detect 44 new NPC2 encoding sequences, raising the total repertoire in these species to 119 (Table 1). It is worth noting, however, that a non-negligible number of these new identified genes are incomplete, likely caused either by a poor genome assembly quality (indicated as the N50 and the number of scaffolds) or a

low number of annotated proteins in the input GFF requiring to predict novel gene models in BITACORA second round (only 42.5% and 63% of the uncovered GR and NPC2 proteins, respectively, were complete; Table 1), demonstrating that the performance of BITACORA depends on both the quality of input annotations and genome assemblies in addition to the specific focal gene family.

Discussion

Gene families are one of the most abundant and dynamic components of eukaryotic genomes. Therefore, having curated genomic data is fundamental not only to carry out comprehensive comparative or functional genomics studies on gene families, but also to understand global genome architecture and biology. During the last decades, the rapid development of sequencing technologies has enabled the rapid accumulation of genome sequences of non-model organisms. Nevertheless, most of them still remain quite fragmented and only have very preliminary and incomplete automatic annotations. The proteins predicted by automatic annotation tools often contain systematic errors, such as incomplete or chimeric gene models, which are especially notable in gene families given the repetitive nature of their members. Besides, since new copies commonly arise by unequal crossing-over, they are frequently found in physically close tandem arrays of similar sequences, further complicating annotations (Clifton et al., 2017; Vieira et al., 2007).

With this in mind, we have developed a bioinformatics tool that helps researchers to access these automatic annotations, extract the information of focal gene families, curate and update gene models and identify new copies from DNA sequences. Using BITACORA, gene family annotations can be really improved using both HMM profiles and iterative searches that incorporate the new variability found in previous searches.

One of the analyses on gene families more sensitive to the quality of annotations is the estimation of the number of gene gains and losses and the associated birth and death rates. The example of the GR family in chelicerates demonstrates the importance of refining annotations using BITACORA. Indeed, using unsupervised annotations in non-model organism genomes directly to estimate turnover rates might produce very erroneous results, not only in terms of gene counts but also in calculations biased to highly expressed and/or very recent copies. Then, BITACORA can be used to reduce considerably these errors and make more accurate and robust inferences about the age/origin of the family and of its mode of evolution.

On the other hand, the curation of both existing and new identified members of a family with BITACORA might be also crucial for further analysis on their sequence evolution. The quality of multiple sequence alignments, which are used to determine orthology groups, to obtain divergence estimates or to detect the footprint of natural selection in gene family members, is strongly compromised by the presence of badly annotated copies, including chimeras and incorrectly annotated fragments. Using BITACORA we can detect these artifacts and either fix or discard them from further analyses.

Despite its proven utility, we are aware that BITACORA do not provide perfect annotations for a gene family. For this reason, we configured the pipeline output to be easily readable for genome editor tools, such as Apollo, which facilitate researchers to improve gene models. Fig. 3 show an example of the annotation tracks generated by BITACORA (BED files) for a member of the candidate carrier protein (*Ccp*: Vizueta et al., 2017) in the genome draft of the spider *Dysdera silvatica* (unpublished data). The automatic annotation using MAKER2 (track GFF3 Dsil) generated an incomplete gene model (with three missing putative exons) that could be easily improved given its identification with BITACORA and the generated output.

Conclusion

Genome annotation, especially in non-model organisms, is still a bottleneck for evolutionary and functional genomic analyses. To assists this task, we developed a comprehensive pipeline that facilitates the curation of existing models and the identification of new gene family copies in genome assemblies with available annotation features and/or genomic sequences. The output of BITACORA can be used as a baseline for manual annotation in genomic annotation editors, used as evidence in automatic annotation tools to improve gene family model predictions, or to directly perform downstream analysis. Future directions should include the implementation in our pipeline of an automatic annotation tool to directly predict new gene models in DNA sequences and its integration as a part of genome annotation editors to facilitate gene family annotation in collaborative genome projects.

Acknowledgements

We would like to thank Paula Escuer and Vadim Pisarenco for helpful discussions. This work was supported by the Ministerio de Economía y Competitividad of Spain (CGL2013-45211, CGL2016-75255) and the Comissió Interdepartamental de Recerca I Innovació Tecnològica of Catalonia, Spain (2017SGR1287). J.V. was supported by a FPI grant (Ministerio de Economía y Competitividad of Spain, BES-2014-068437).

Author contributions

J.V., A.S.-G and J.R. conceived the work. J.V. wrote the scripts, did the analyses and wrote the first version of the manuscript. All authors checked and confirmed the final version of the manuscript.

Data accessibility

BITACORA is available from http://www.ub.edu/softevol/bitacora, and https://github.com/molevol-ub/bitacora

References

Altschul, S. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. doi:10.1093/nar/25.17.3389

Clifton, B. D., Librado, P., Yeh, S.-D., Solares, E. S., Real, D. A., Jayasekera, S. U., … Ranz, J. M. (2017). Rapid Functional and Sequence Differentiation of a Tandemly Repeated Species-Specific Multigene Family in *Drosophila*. *Molecular Biology and Evolution*, *34*(1), 51–65. doi:10.1093/molbev/msw212

Dominguez Del Angel, V., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Vinnere Pettersson, O., … Lantz, H. (2018). Ten steps to get started in Genome Assembly and Annotation. *F1000Research*, *7*, ELIXIR-148. doi:10.12688/f1000research.13598.1

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, *7*(10), e1002195. doi:10.1371/journal.pcbi.1002195

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., … Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Research*, *42*(Database issue), D222–D230. doi:10.1093/nar/gkt1223

Gremme, G., Brendel, V., Sparks, M. E., & Kurtz, S. (2005). Engineering a software tool for gene structure prediction in higher organisms. *Information and Software Technology*, *47*(15), 965–978. doi:10.1016/J.INFSOF.2005.09.005

Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., & Stanke, M. (2016). BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics*, *32*(5), 767–769. doi:10.1093/bioinformatics/btv661

Holt, C., & Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, *12*(1), 491. doi:10.1186/1471-2105-12-491

Joseph, R. M., & Carlson, J. R. (2015). *Drosophila* Chemoreceptors: A Molecular Interface Between the Chemical World and the Brain. *Trends in Genetics : TIG*, *31*(12), 683–695. doi:10.1016/j.tig.2015.09.005

Keller, O., Kollmar, M., Stanke, M., & Waack, S. (2011). A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics*, *27*(6), 757–763. doi:10.1093/bioinformatics/btr010

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, *5*, 59. doi:10.1186/1471-2105-5-59

Lee, E., Helt, G. A., Reese, J. T., Munoz-Torres, M. C., Childers, C. P., Buels, R. M., … Lewis, S. E. (2013). Web Apollo: a web-based genomic annotation editing platform. *Genome Biology*, *14*(8), R93. doi:10.1186/gb-2013-14-8-r93

Lomsadze, A., Burns, P. D., & Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research*, *42*(15), e119–e119. doi:10.1093/nar/gku557

Pelosi, P., Iovinella, I., Felicioli, A., & Dani, F. R. (2014). Soluble proteins of chemical communication: an overview across arthropods. *Frontiers in Physiology*, *5*(August), 320. doi:10.3389/fphys.2014.00320

Robertson, H. M. (2015). The Insect Chemoreceptor Superfamily Is Ancient in Animals. *Chemical Senses*, *40*(9), 609–614. doi:10.1093/chemse/bjv046

Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, *6*, 31. doi:10.1186/1471-2105-6-31

Stanke, M., & Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, *19*(Suppl 2), ii215–ii225. doi:10.1093/bioinformatics/btg1080

Stanke, Mario, Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, *24*(5), 637–644. doi:10.1093/bioinformatics/btn013

Vieira, F. G., Sánchez-Gracia, A., & Rozas, J. (2007). Comparative genomic analysis of the odorant-binding protein family in 12 *Drosophila* genomes: purifying selection and birth-and-death evolution. *Genome Biology*, *8*(11), R235. doi:10.1186/gb-2007-8-11-r235

Vizueta, J., Frías-López, C., Macías-Hernández, N., Arnedo, M. A., Sánchez-Gracia, A., & Rozas, J. (2017). Evolution of chemosensory gene families in arthropods: Insight from the first inclusive comparative transcriptome analysis across spider appendages. *Genome Biology and Evolution*, *9*(1), 178–196. doi:10.1093/gbe/evw296

Vizueta, J., Rozas, J., & Sánchez-Gracia, A. (2018). Comparative Genomics Reveals Thousands of Novel Chemosensory Genes and Massive Changes in Chemoreceptor Repertories across Chelicerates. *Genome Biology and Evolution*, *10*(5), 1221–1236. doi:10.1093/gbe/evy081

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, *13*(5), 329–342. doi:10.1038/nrg3174

Yohe, L. R., Davies, K. T. J., Simmons, N. B., Sears, K. E., Dumont, E. R., Rossiter, S. J., & Dávalos, L. M. (2019). Evaluating the performance of targeted sequence capture, RNA-Seq, and degenerate-primer PCR cloning for sequencing the largest mammalian multigene family. *Molecular Ecology Resources*. doi:10.1111/1755-0998.13093

Zhang, Y., Zheng, Y., Li, D., & Fan, Y. (2014). Transcriptomics and identification of the chemoreceptor superfamily of the pupal parasitoid of the oriental fruit fly, *Spalangia endius* Walker (Hymenoptera: Pteromalidae). *PloS One*, *9*(2), e87800. doi:10.1371/journal.pone.0087800

Tables

Table 1. Summary of the number of GRs and NPC2 genes identified by BITACORA in seven chelicerate genomes.

Figures

Fig. 1. BITACORA workflow.

Fig. 2. Phylogenetic relationships among the seven chelicerate species surveyed for the GR and the NPC2 families.

Figure 3. Visualization in Apollo genome editor of the BITACORA output with the annotation features of a candidate carrier protein (*Ccp*) gene (Dsil_g69.t1 in *D. silvatica*). The Dsil GFF3 track shows the original GFF file obtained by BRAKER in the focal genome. The two BED tracks shows the output files generated in BITACORA showing three putative exons identified in a region not annotated in the original GFF and all the six exons identified by BITACORA. The RNA track (PalpRNA_bw) shows the genomic regions with mapped reads from the sequencing of the spider palp RNA-seq library. The final gene model is shown in the User-created Annotations track.

Supplementary Material

BITACORA Documentation

Table 1. Summary of the number of GRs and NPC2 genes identified by BITACORA in seven chelicerate genomes

Species	Genome Information[a]			GR				NPC2			
				BITACORA identified sequences[b]	Identified in		Complete sequences[e]	BITACORA identified sequences[b]	Identified in		Complete sequences[e]
	N50	Scaffolds	Predicted proteins		GFF[c]	Genome[d]			GFF[c]	Genome[d]	
C. sculpturatus	342,549	10,457	30,465	1,234	24	1,210	518	15	11	4	7
M. martensii	45,228	92,408	32,016	673	23	650	173	15	9	6	7
A. geniculata	20,294	4,986,575	76,238	137	2	135	54	20	14	6	13
Lo. reclusa	63,233	143,678	20,617	203	1	202	148	18	7	11	9
S. mimosarum	480,636	68,653	26,888	229	4	225	120	16	13	3	14
P. tepidariorum	465,572	31,445	32,186	819	52	767	383	20	19	1	17
La. hesperus	13,889	151,814	17,364	76	0	76	37	15	2	13	8
Total chelicerata	-	-	-	3,371	106	3,265	1,433	119	75	44	75

[a] Features from the focal genomes

[b] Total number of members of the gene family after running the two steps of BITACORA

[c] Total number of members of the gene family predicted in the former GFF

[d] Total number of new gene family members identified by BITACORA (output of Step 2)

[e] Number of complete sequences identified by BITACORA

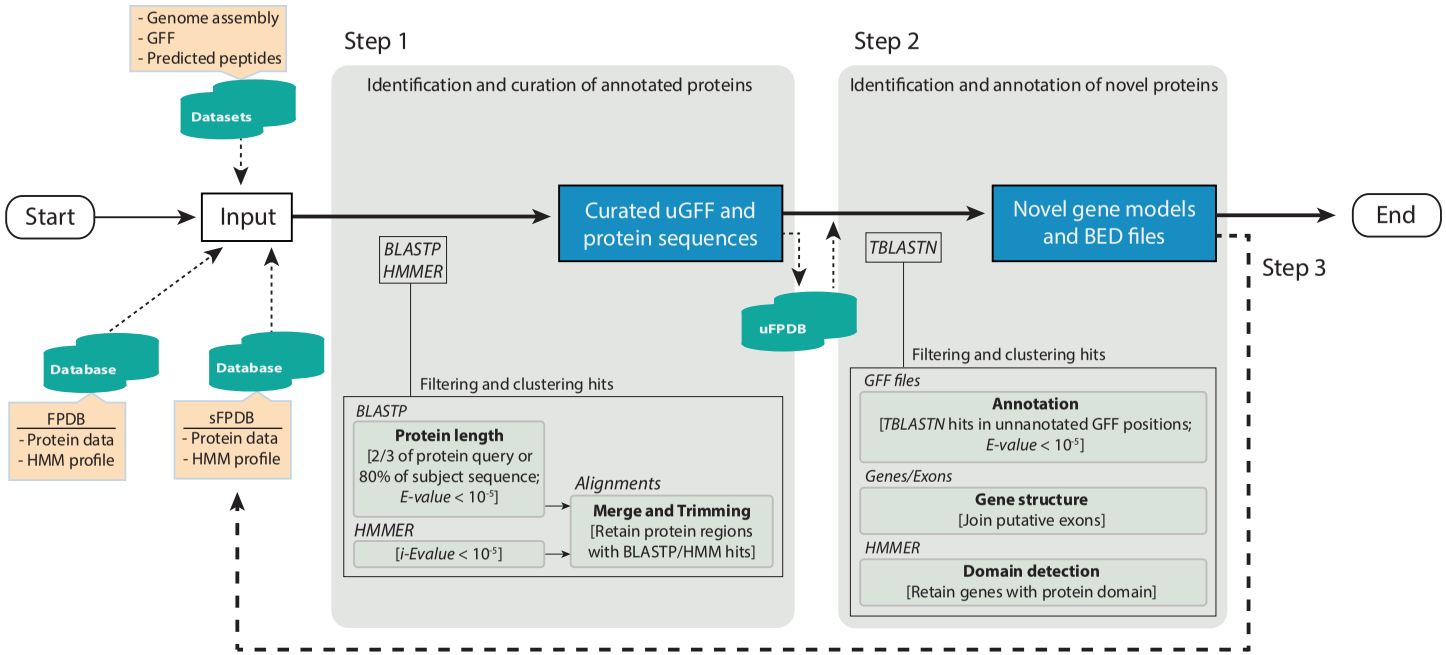


Figure 1. BITACORA workflow

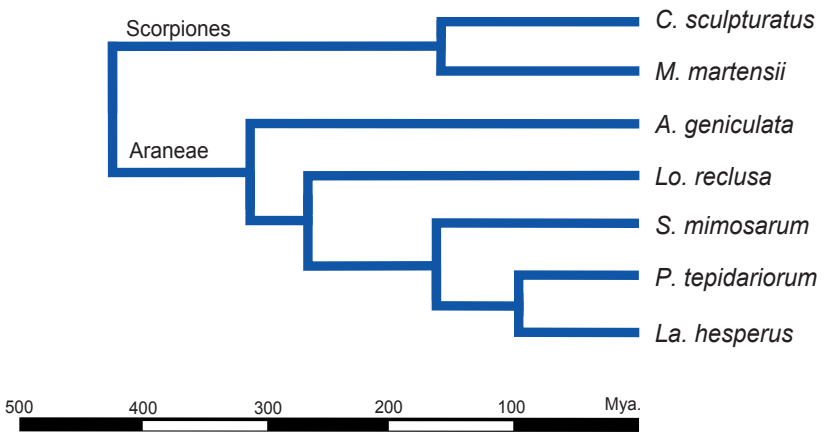


Figure 2. Phylogenetic relationships among the seven chelicerate species surveyed for the GR and the NPC2 families.

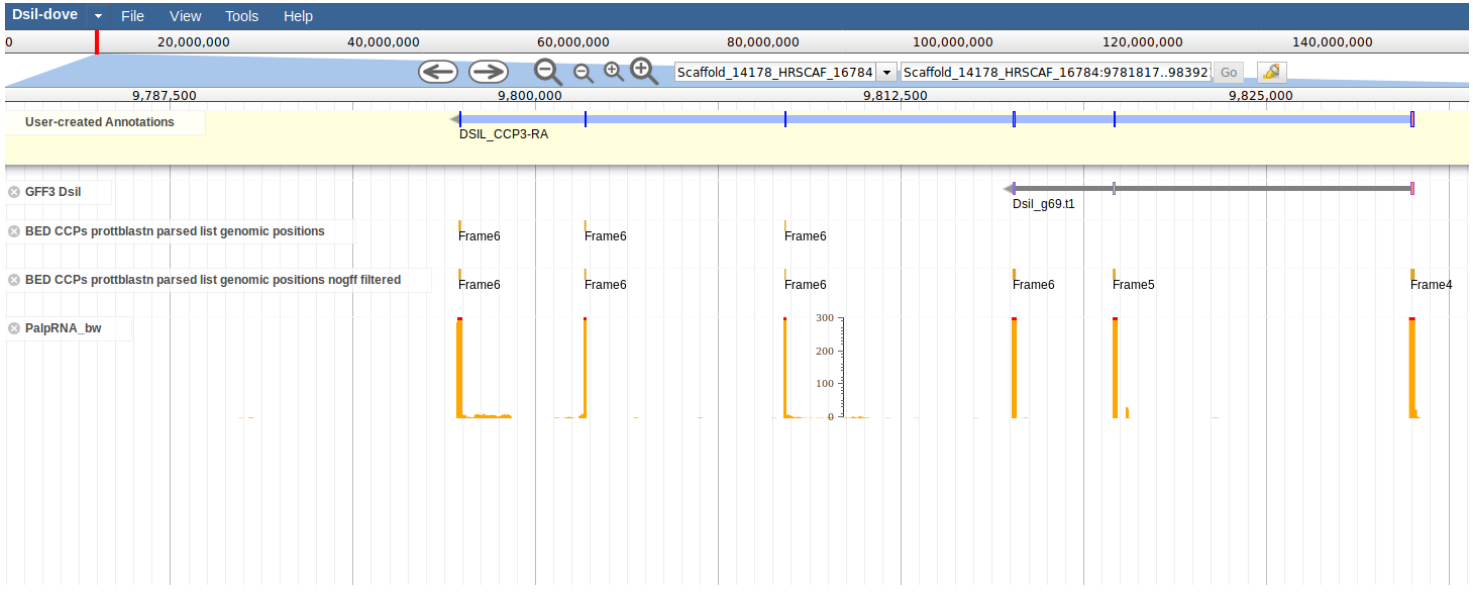


Figure 3. Visualization in Apollo genome editor of the BITACORA output with the annotation features of a candidate carrier protein (*Ccp*) gene (Dsil_g69.t1 in *D. silvatica*). The Dsil GFF3 track shows the original GFF file obtained by BRAKER in the focal genome. The two BED tracks shows the output files generated in BITACORA showing three putative exons identified in a region not annotated in the original GFF and all the six exons identified by BITACORA. The RNA track (PalpRNA_bw) shows the genomic regions with mapped reads from the sequencing of the spider palp RNA-seq library. The final gene model is shown in the User-created Annotations track.

BITACORA: A comprehensive tool for the identification and annotation of gene families in genome assemblies

Vizueta J., Sánchez-Gracia A. and Rozas J.

Supplementary Material

BITACORA:

A comprehensive tool for the identification and annotation of gene families in genome assemblies

Joel Vizueta
Alejandro Sánchez-Gracia
Julio Rozas

Departament de Genètica, Microbiologia i Estadística
Institut de Recerca de la Biodiversitat (IRBio)

Universitat de Barcelona

http://www.ub.es/softevol/bitacora

September 17th, 2019

- Overview

Genome annotation is a critical bottleneck in genomic research, especially for the rigorous and comprehensive study of gene families in the genomes of non-model organisms. Despite the recent progress in automatic annotation, the tools developed for this task often produce inaccurate annotations, such as fused, chimeric, partial or even completely absent gene models for many family copies, which require considerable extra efforts to be amended. Here we present BITACORA, a bioinformatics tool that integrates sequence similarity search algorithms and Perl scripts to facilitate both the curation of these inaccurate annotations and the identification of previously undetected gene family copies directly from DNA sequences. The pipeline generates general feature format (GFF) files with both curated and novel gene models, and FASTA files with the predicted proteins. The output of BITACORA can be easily integrated in genomic annotation editors, greatly facilitating subsequent semi-automatic annotation and downstream evolutionary analyses.

Authors

Joel Vizueta	jvizueta@ub.edu
Alejandro Sánchez-Gracia	elsanchez@ub.edu
Julio Rozas	jrozas@ub.edu

BITACORA Publication

Vizueta, J., Sánchez-Gracia, A. and Rozas, J. 2019. BITACORA: A comprehensive tool for the identification and annotation of gene families in genome assemblies. *bioRxiv*. https://doi.org/10.1101/593889.

BITACORA Web Site

www.ub.edu/softevol/bitacora

0 Workflow & Contents

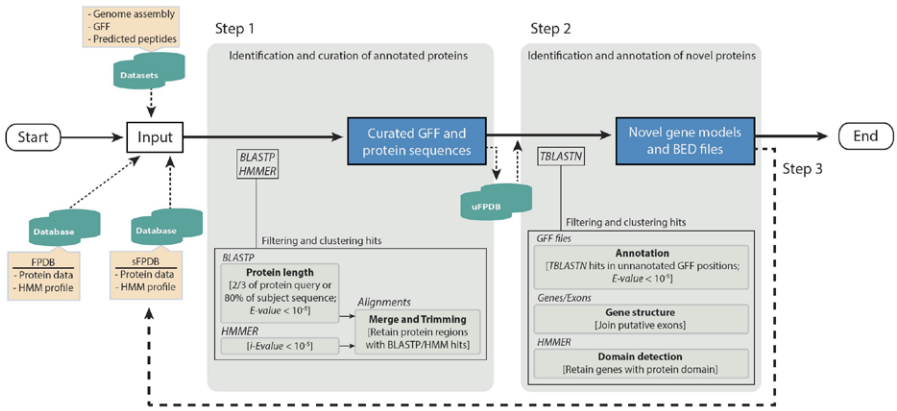


Figure. Workflow showing the basic steps used in BITACORA

1. Installation

2. Prerequisites

3. Computational Requirements

4. Usage modes

- **4.1.** Full mode
- **4.2.** Protein mode
- **4.3.** Genome mode

5. Parameters

6. Running BITACORA

7. Output

8. Example

9. Citation

10. Troubleshooting

1 Installation

BITACORA is distributed as a multiplatform shell script (runBITACORA.sh) that calls several other perl scripts, which include all functions responsible of performing all pipeline tasks. Hence, it does not require any installation or compilation step.

You can download all package contents from GitHub: https://github.com/molevol-ub/bitacora

To run the pipeline edit the master script runBITACORA.sh variables described in Prerequisites, Data, and Parameters.

2 Prerequisites

- BLAST: Download blast executables from:
ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/

- HMMER: The easiest way to install HMMER in your system is to type one of the following commands in your terminal:

```
% brew install hmmer                # OS/X, HomeBrew
% port install hmmer                # OS/X, MacPorts
% apt install hmmer                 # Linux (Ubuntu, Debian...)
% dnf install hmmer                 # Linux (Fedora)
% yum install hmmer                 # Linux (older Fedora)
% conda install -c bioconda hmmer   # Anaconda
```

Or compile HMMER binaries from the source code: http://hmmer.org/

- Perl: Perl is installed by default in most operating systems. See https://learn.perl.org/installing/ for installation instructions.

HMMER and BLAST binaries require to be added to the PATH environment variable. Specify the correct path to bin folders in the master script `runBITACORA.sh`, if necessary.

```
$ export PATH=$PATH:/path/to/blast/bin
$ export PATH=$PATH:/path/to/hmmer/bin
```

3 Computational requirements

BITACORA have been tested in UNIX-based platforms (both in Mac OS and Linux operating systems). Multiple threading can be set in blast searches, which is the most time-consuming step, by editing the option THREADS in `runBITACORA.sh`

For a typical good quality genome (~2Gb in size and ~10,000 scaffolds) and a standard modern PC (16Gb RAM), a full run of BITACORA is completed in less than 24h. This running time, however, will depend on the size of the gene family or the group of genes surveyed in a particular analysis. For gene families of 10 to 100 members, BITACORA spends from minutes to a couple of hours.

In case of larger or very fragmented genomes, BITACORA should be used in a computer cluster or workstation given the increase of RAM memory and time required.

4 Usage modes

4.1. Full mode

BITACORA has been initially designed to work with genome sequences and protein annotations (full mode). However, the pipeline can also be used either with only protein or only genomic sequences (protein and genome modes, respectively). These last modes are explained in next subsections.

Preparing the data: The input files (in plain text) required by BITACORA to run a full analysis are (update the complete path to these files in the master script `runBITACORA.sh`):

I. File with genomic sequences in FASTA format

II. File with structural annotations in GFF3 format. [*NOTE: mRNA* or *transcript*, and *CDS* are mandatory fields].

---------- GFF3 example

```
lg1_ord1_scaf1770	AUGUSTUS	gene	13591	13902	0.57	+	.	ID=g1;
lg1_ord1_scaf1770	AUGUSTUS	mRNA	13591	13902	0.57	+	.	ID=g1.t1;Parent=g1;
lg1_ord1_scaf1770	AUGUSTUS	start_codon	13591	13593	.	+	0	Parent=g1.t1;
lg1_ord1_scaf1770	AUGUSTUS	CDS	13591	13902	0.57	+	0	ID=g1.t1.CDS1;Parent=g1.t1
lg1_ord1_scaf1770	AUGUSTUS	exon	13591	13902	.	+	.	ID=g1.t1.exon1;Parent=g1.t1;
lg1_ord1_scaf1770	AUGUSTUS	stop_codon	13900	13902	.	+	0	Parent=g1.t1;
```

BITACORA also accepts other GFF formats, such as Ensembl GFF3 or GTF. [NOTE: GFF formatted files from NCBI can cause errors when processing the data, use the supplied script "`reformat_ncbi_gff.pl`" (located in the folder `/Scripts/Tools`) to make the file parsable by BITACORA]. See Troubleshooting in case of getting errors while parsing your GFF.

---------- Ensembl GFF3 example

```
AFFK01002511  EnsemblGenomes  gene        761  1018  .  -  .  ID=gene:SMAR013822;assembly_name=Smar1;biotype=protein_coding;logic_name=ensemblgenomes;versio
AFFK01002511  EnsemblGenomes  transcript  761  1018  .  -  .  ID=transcript:SMAR013822-RA;Parent=gene:SMAR013822;assembly_name=Smar1;biotype=protein
AFFK01002511  EnsemblGenomes  CDS         761  811   .  -  0  Parent=transcript:SMAR013822-RA;assembly_name=Smar1
AFFK01002511  EnsemblGenomes  exon        761  811   .  -  .  Parent=transcript:SMAR013822-RA;Name=SMAR013822-RA-E2;assembly_name=Smar1;constitutive=1;ensem
AFFK01002511  EnsemblGenomes  CDS         887  1018  .  -  0  Parent=transcript:SMAR013822-RA;assembly_name=Smar1
AFFK01002511  EnsemblGenomes  exon        887  1018  .  -  .  Parent=transcript:SMAR013822-RA;Name=SMAR013822-RA-E1;assembly_name=Smar1;constitutive=1;ensem
```

III. File with predicted proteins in FASTA format. BITACORA requires identical IDs for proteins and their corresponding mRNAs or transcripts IDs in the GFF3. [NOTE: we recommend using genes but not isoforms in BITACORA; isoforms can be removed or properly annotated after BITACORA analysis]

IV. Specific folder with files containing the query protein databases (`YOURFPDB_db.fasta`) and HMM profiles (`YOURFPDB_db.hmm`) in FASTA and hmm format, respectively, where the "YOURFPDB" label is your specific data file name. The addition of "`_db`" to the database name with its proper extension, `fasta` or `hmm`, is mandatory.
BITACORA requires one protein database and profile per surveyed gene family (or gene group). See `Example/DB` files for an example of searching for two different gene families in BITACORA: OR, Odorant Receptors; and CD36-SNMP.

[NOTE: profiles covering only partially the proteins of interest are not recommended]

Notes on HMM profiles:

HMM profiles are found in InterPro or PFAM databases associated to known protein domains. If you don't know if your protein contains any described domain, you can search in InterPro (http://www.ebi.ac.uk/interpro/) using the protein sequence of one of your queries to identify domains.

For example, for the chemosensory proteins (CSPs) in insects, you can download the HMM profile from pfam (Curation & model PFAM submenu):

http://pfam.xfam.org/family/PF03392#tabview=tab6

In the case of searching for proteins with not described protein domains, or with domains not covering most of the protein sequence, it should be performed an alignment of the query proteins to create a specific HMM profile.

Example of building a protein profile (it requires an aligner, here we use mafft as example):

```
$ mafft --auto FPDB_db.fasta > FPDB_db.aln
$ hmmbuild FPDB_db.hmm FPDB_db.aln
```

Notes on the importance of selecting a confident curated database:

The proteins included in the database to be used as query (FPDB) in the protein search is really important; indeed, the inclusion of unrelated or bad annotated proteins could lead to the identification and annotation of proteins unrelated to the focal gene family and can inflate the number of sequences identified.

On the other hand, if possible, we recommend to include proteins from phylogenetically-close species to increase the power of identifying proteins, particularly in fast-evolving and divergent gene families. If your organism of interest does not have an annotated genome of a close related species, we suggest to perform a second BITACORA round (step 3 described in the manuscript), including in the query database (sFPDB) the sequences identified in the first round, along with a new HMM profile build with these sequences. This step may facilitate the identification of previously undetected related divergent sequences.

4.2. Protein mode

BITACORA can also run with a set of proteins (i.e. predicted proteins from transcriptomic data; script `runBITACORA_protein_mode.sh`) by using the input files described in points **III** and **IV** of the section **4.1**.

Under this mode, BITACORA identifies, curates when necessary, and report all members of the surveyed family among the predicted proteins. The original protein sequences (not being curated) are also reported (located in `Intermediate_Files` if cleaning output is active).

4.3. Genome mode

BITACORA can also run with raw genome sequences (i.e., not annotated genomes; script `runBITACORA_genome_mode.sh`), by using the input files described in points **I** and **IV** of the section **4.1**.

Under this mode, BITACORA identifies *de novo* all members of the surveyed family and returns a BED file with gene coordinates of the detected exons, a FASTA file with predicted proteins from these exons and a GFF3 file with the corresponding structural annotations.

[NOTE: The gene models generated under this mode are only semi-automatic predictions and require further manual annotation, i.e. using genomic annotation editors, such as Apollo. The output file of the genome mode can also be used as protein evidence in automatic annotators as MAKER2 or BRAKER1 (see output section)]

5 Parameters

- The option CLEAN can be used to create the `Intermediate_files` directory where all intermediate files will be stored (see output section).

```
CLEAN=T  #T=true, F=false
```

- BLAST and HMMER hits are filtered with a default cut-off E-value of 10e-5 (in addition to an internal parameter for filtering the length covered by the alignment).

E-value can be modified in the master script runBITACORA.sh:

```
EVALUE=10e-5
```

- Number of threads to be used in blast searches, default is 1.

```
THREADS=1
```

- BITACORA uses by default a value of 15 kb to join putative exons from separate but contiguous (and in the same scaffold) genome hits.

This value can be modified in the master script runBITACORA.sh:

```
MAXINTRON=15000
```

Notes on the parameter MAXINTRON:

Estimating the intron length distribution in your genome:

MAXINTRON is a critical parameter affecting the quality of the gene models built after joining *de novo* identified exons after BLASTN search (see BITACORA article). BITACORA is distributed with a script (`get_intron_size_fromgff.pl`) to compute some summary statistics, such as the mean, median, and the 95% and 99% upper limits of the intron length distribution, of an input GFF, which can contain all genes from genome or only the genes identified for a particular gene family (i.e. GFF generated in BITACORA output).

Note that a very high value could join exons from different genes, generating a putative chimeric gene. On the other hand, a very low value could not join exons from the same gene. Therefore, it is very important to set a MAXINTRON biological realistic value, which could vary across species or assemblies. As default, BITACORA uses a conservative high value, as a compromise between ensuring the joining of all exons from a same gene, and avoiding the generation of erroneous gene fusions. In any case, a large value of MAXINTRON parameter prevents the annotation of fragmented genes but can generate gene models with multiple gene fusions. Putative gene fusions (proteins with two or more domains predicted by BITACORA) are tagged with the label "`Xdom`" at the end of the protein name in the output file, being X the number of putative genes (detected domains).

The number of putative fussed genes identified as new proteins in not annotated regions of the genome can be obtained using the following command in the terminal:

```
$ grep '>.*dom' DB/DB_genomic_and_annotated_proteins_trimmed.fasta
```

6 Running BITACORA

After preparing the data as indicated in steps 5 (Usage) and 6 (Parameters), you can execute BITACORA with the following command:

```
$ bash runBITACORA.sh
```

7 Output

BITACORA creates an output folder for each query database, and three files with the number of proteins identified in each step, including a summary table. For the genome and protein modes, only one summary table will be reported with the number of identified genes.

In each folder, there are the following **main files** (considering you chose to clean output directory. If not, all files will be found in the same output folder):

- `YOURFPDB_genomic_and_annotated_genes_trimmed.gff3`: GFF3 file with information of all identified protein curated models both in already annotated proteins and unannotated genomic sequences.

- `YOURFPDB_genomic_and_annotated_proteins_trimmed.fasta`: A fasta file containing the protein sequences from the above gene models.

Non-redundant data: Relevant information excluding identical proteins, or those considered as artefactual false positives (i.e. duplicated scaffolds, isoforms…).

- `YOURFPDB_genomic_and_annotated_genes_trimmed_nr.gff3`: GFF3 file containing all identified non-redundant protein curated models both in already annotated proteins and unannotated genomic sequences.

- `YOURFPDB_genomic_and_annotated_proteins_trimmed.fasta`: A fasta file containing the non-redundant protein sequences from the above gene models.

BED files with non-redundant merged blast hits in genome sequence:

- `YOURFPDBtblastn_parsed_list_genomic_positions.bed`: BED file with only merged blast alignments in non-annotated regions.

- `YOURFPDBtblastn_parsed_list_genomic_positions_nogff_filtered.bed`: BED file with merged blast alignment in all genomic regions.

In addition, BITACORA generates the following **Intermediate files** (located into Intermediate_files folder created in cleaning, if active):

- `YOURFPDB_annot_genes.gff3` and `YOURFPDB_proteins.fasta`: GFF3 and fasta file containing the original untrimmed models for the identified proteins.

- YOURFPDB_annot_genes_trimmed.gff3 and YOURFPDB_proteins_trimmed.fasta: GFF3 and fasta containing only the curated model for the identified annotated proteins (trimming exons if not aligned to query FPDB sequences or split putative fused genes)

- YOURFPDB_genomic_genes.gff3: GFF3 containing novel identified proteins in genomic sequences.

- YOURFPDB_genomic_genes_trimmed.gff3: GFF3 containing novel identified proteins in genomic sequences curated by the positions identified in the HMM profile.

- YOURFPDBgfftrimmed.cds.fasta and YOURFPDBgfftrimmed.pepfasta: Files containing CDS and protein sequences translated directly from YOURFPDB_annot_genes_trimmed.gff3

- YOURFPDBgffgenomictrimmed.cds.fasta and YOURFPDBgffgenomictrimmed.pep.fasta: Files containing CDS and protein sequences translated directly from YOURFPDB_genomic_genes_trimmed.gff3

- hmmer folder containing the output of HMMER searches against the annotated proteins and novel proteins identified in the genome

- YOURFPDB_blastp.outfmt6: BLASTP output of the search of the query FPDB against the annotated proteins

- YOURFPDB_tblastn.outfmt6: TBLASTN output of the search of the query FPDB against the genomic sequence

- YOURFPDB_blastp_parsed_list.txt; YOURFPDB_hmmer_parsed_list.txt; YOURFPDB_allsearches_list.txt; YOURFPDB_combinedsearches_list.txt: Parsed files combining all hits and extending the hit positions from BLASTP and HMMER outputs

- YOURFPDB_tblastn_parsed_list_genomic_positions.txt (and _notgff_filtered) : File containing the positions identified after parsing the TBLASTN search.

- YOURFPDB_prots_VsGFF_badannot_list.txt and _goodannot_list.txt: Debugging files: These files are for checking that the identified proteins and the protein models in the GFF3 codify the same protein. If the file badannot_list.txt contains some identifier, it means that the GFF3 annotation is incorrect pointing to a bad annotation in the original GFF3. Please, try to translate the CDS for that protein into the 3 reading frames and check if the 2nd or 3rd frame codify for the protein in question stored in "YOURFPDB_genomic_and_annotated_proteins_trimmed_nr.fasta". If correct, modify the GFF3 by adding 1 or 2 nucleotide position in the start of the GFF3 (take into account if it is transcribed from forward or reverse strand). If negative, please report the error via GitHub.

- YOURFPDB_genomic_genes_proteins.fasta: It contains all merged exons from putative novel proteins identified in the genome before filtering those without the protein domain identified with HMMER.

- YOURFPDB_genomic_exon_proteins.fasta: contains the exon sequences joined into genes in the aforementioned file.

- Additional generated files are stored for pipeline debugging and controls.

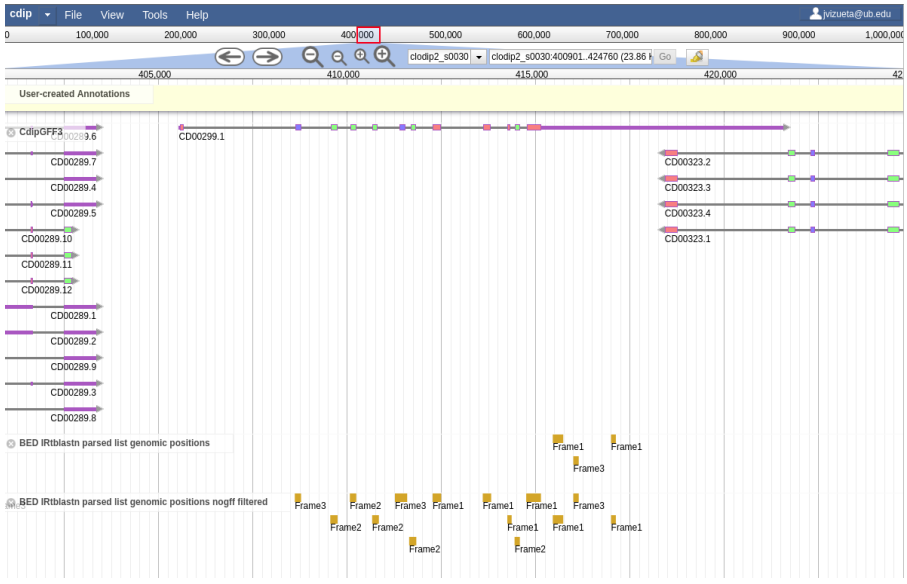
Notes on BITACORA output:

The obtained proteins could be used for further prospective analyses or to facilitate a more curated annotation using genome annotation editors or, in the case of having a high number of not annotated proteins in the GFF, BITACORA output sequences could also be used as evidence to improve the annotation of automatic annotators as MAKER2 or BRAKER1. However, a first validation of the obtained proteins should be performed, more specifically in those obtained newly from genome (taking into account the parameter used to join putative exons, to split putative joined genes or join exons from the same gene). In addition, these proteins obtained and assembled from genomic regions are illustrative, but more putative genes (true negatives) could be obtained from the TBLASTN BED file positions discarded for not being identified with the protein domain (i.e. alignments containing introns between two proximal exons could lead not to identify the domain in the protein).

Such validation to identify putative erroneously assigned proteins (mainly caused by the inclusion of contaminant sequences in the query database) could consist in aligning all proteins and checking the MSA, constructing the phylogeny of the gene with related species or the gene family; doing a reduced blast with NCBI-nr database or obtaining structural particularities of the proteins (i.e. characterizing protein domains as transmembrane domains, signal peptides...). See our manuscript Vizueta et al. (2018) for an example of such analyses.

In particular, BITACORA full and genome mode is also designed to facilitate the gene annotation in editors as Apollo. For that, the use of the following files would be useful (see an example in Documentation/example_Apollo.png):

- Original GFF3
- Final GFF3 with curated models for the annotated proteins
- BED file from TBLASTN search



If there are sequences containing stop codons, codified as "X", it could be artefactual from TBLASTN hits if they are in the beginning or end of an exon or, otherwise, those genes are probably pseudogenes.

Nonetheless, again, new proteins identified from unannotated genomic regions should be properly annotated using genome browser annotation tools such as Apollo, or could be used as evidence to improve the annotation of automatic annotators as MAKER2 or BRAKER1. We estimate an approximate number of them which could be used for prospective analyses.

8 Example

An example to run BITACORA can be found in `Example` folder. First, unzip the `Example_files.zip` file to obtain the necessary files for BITACORA. In this example, two chemosensory-related gene families in insects: Odorant receptors (ORs), and the CD36-SNMP gene family; will be searched in the chromosome 2R of *Drosophila melanogaster*. The GFF3 and protein files are modified from original annotations, deleting some gene models, to allow that BITACORA can identify novel not-annotated genes.

To run the example, edit the master script `runBITACORA.sh` to add the path to BLAST and HMMER binaries and run the script. It will take around 1 minute with 2 threads.

```
$ bash runBITACORA.sh
```

9 Citation

Joel Vizueta, Alejandro Sánchez-Gracia, and Julio Rozas. 2019. BITACORA: A comprehensive tool for the identification and annotation of gene families in genome assemblies. *bioRxiv*. https://doi.org/10.1101/593889

Moreover, you can also cite the following article where we describe the protein annotation procedure:

Joel Vizueta, Julio Rozas, Alejandro Sánchez-Gracia; Comparative Genomics Reveals Thousands of Novel Chemosensory Genes and Massive Changes in Chemoreceptor Repertories across Chelicerates, Genome Biology and Evolution, Volume 10, Issue 5, 1 May 2018, Pages 1221–1236, https://doi.org/10.1093/gbe/evy081

10 Troubleshooting

When BITACORA detects any error related to input data, it stops and prints the description of the error. Please check the error and your data.

If you are getting errors related to parsing the GFF file, take into account that BITACORA expects proteins ID to be as ID in mRNA rows from GFF3.

In case of protein ID and mRNA ID causing error as they are not the named equally, first, you can use the script located in `Scripts/Tools/get_proteins_notfound_ingff.pl` to check which proteins are not found in the GFF3 file, as detailed in the Error message. You could use only those proteins found in the GFF3 in BITACORA.

If all proteins are named differently in the GFF3, you can obtain a protein file from the GFF3 using the script `Scripts/gff2fasta_v3.pl` and use that protein file as input to BITACORA.

You could also modify the perl module `Readgff.pm` to allow BITACORA to read your data. Otherwise, modify the GFF, preferably, as GFF3 format.

If you cannot solve the error, create an issue in Github specifying the error and all details as possible.

2

Evolution of chemosensory gene families in arthropods: Insight from the first inclusive comparative transcriptome analysis across spider appendages

A diferencia de hexápodos y vertebrados, el conocimiento de las moléculas específicas involucradas en la quimiopercepción en quelicerados procede únicamente del análisis comparativo de secuencias genómicas. Estos análisis han revelado que los genomas de ácaros y arañas contienen genes que codifican secuencias homólogas de algunos receptores y proteínas solubles identificadas en el sistema quimiosensorial de insectos. En este estudio hemos llevado a cabo el primer análisis comparativo de RNA-seq en distintas estructuras corporales de un quelicerado: la araña cazadora nocturna *Dysdera silvatica* Schmidt 1981. En particular, hemos obtenido el transcriptoma completo de esta especie, así como los perfiles de expresión específicos en los palpos y el primer par de patas, descritos como aprendices olfativos en arañas, y el resto de patas, las que también tienen pelos que han sido identificados morfológicamente como quimiosensoriales. Hemos identificado receptores ionótropicos (IR) y gustativos (GR) que se expresan de forma específica o diferencial entre los distintos tejidos, algunos de ellos en los apéndices qumiosensoriales. Además, estos IRs son los únicos receptores conocidos con expresión en estas estructuras que tienen una función olfativa. Los resultados, integrados con un extenso análisis filogenético en artrópodos, muestran una expresión diferencial de los genes quimiosensoriales en el cuerpo de *D. silvatica*, y sugieren que algunos IRs podrían mediar la señal olfativa en quelicerados. A su vez, hemos detectado la expresión de una familia multigénica relacionada con los OBPs (*odorant-binding proteins*) de insectos, lo que sugiere que esta familia es más antigua de lo que se creía, así como la identificación de una familia multigénica no caracterizada, expresada en los apéndices quimiosensoriales, que codifica pequeñas proteínas globulares de unión a ligando, y que podría ser una buena candidata a participar en la quimiopercepción.

GBE

Evolution of Chemosensory Gene Families in Arthropods: Insight from the First Inclusive Comparative Transcriptome Analysis across Spider Appendages

Joel Vizueta[1], Cristina Frías-López[1], Nuria Macías-Hernández[2], Miquel A. Arnedo[2], Alejandro Sánchez-Gracia[1,*], and Julio Rozas[1,*]

[1]Departament de Genètica, Microbiologia i Estadística and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Spain

[2]Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Spain

*Corresponding authors: E-mails: jrozas@ub.edu; elsanchez@ub.edu.

Accepted: December 16, 2016

Data deposition: This project has been deposited at the Sequence Read Archive (SRA) database under accession numbers SRX1612801, SRX1612802, SRX1612803 and SRX1612804 (Bioproject number: PRJNA313901).

Abstract

Unlike hexapods and vertebrates, in chelicerates, knowledge of the specific molecules involved in chemoreception comes exclusively from the comparative analysis of genome sequences. Indeed, the genomes of mites, ticks and spiders contain several genes encoding homologs of some insect membrane receptors and small soluble chemosensory proteins. Here, we conducted for the first time a comprehensive comparative RNA-Seq analysis across different body structures of a chelicerate: the nocturnal wandering hunter spider *Dysdera silvatica* Schmidt 1981. Specifically, we obtained the complete transcriptome of this species as well as the specific expression profile in the first pair of legs and the palps, which are thought to be the specific olfactory appendages in spiders, and in the remaining legs, which also have hairs that have been morphologically identified as chemosensory. We identified several ionotropic (*Ir*) and gustatory (*Gr*) receptor family members exclusively or differentially expressed across transcriptomes, some exhibiting a distinctive pattern in the putative olfactory appendages. Furthermore, these IRs were the only known olfactory receptors identified in such structures. These results, integrated with an extensive phylogenetic analysis across arthropods, uncover a specialization of the chemosensory gene repertoire across the body of *D. silvatica* and suggest that some IRs likely mediate olfactory signaling in chelicerates. Noticeably, we detected the expression of a gene family distantly related to insect odorant-binding proteins (OBPs), suggesting that this gene family is more ancient than previously believed, as well as the expression of an uncharacterized gene family encoding small globular secreted proteins, which appears to be a good chemosensory gene family candidate.

Key words: chemosensory gene families, specific RNA-Seq, *de novo* transcriptome assembly, functional annotation, chelicerates, arthropods.

Introduction

Chemoreception, the detection and processing of chemical signals in the environment, is a biological process that is critical for animal survival and reproduction. The essential role of smell and taste in the detection of food, hosts and predators and their participation in social communication make the molecular components of this system solid candidates for important adaptive changes associated with animal terrestrialization (Whiteman and Pierce 2008). In insects, chemical recognition occurs in specialized hair-like cuticular structures called sensilla, which can be found almost anywhere in the body (Joseph and Carlson 2015). In *Drosophila*, olfactory sensilla are concentrated on the antenna and the maxillary palps, while gustatory sensilla are spread across various body locations, such as the proboscis, the legs and the anterior margins of wings (Pelosi 1996; Shanbhag et al. 2001). The chemoreceptor proteins embedded within the membrane of sensory neurons (SN) innervating these sensilla are responsible for transducing the external chemical signal into an action potential. In the case of smell, olfactory SNs project the axons to

specific centers of the brain, where the signals are processed and engender a behavioral response to the specific external stimuli. The process can be facilitated by small soluble chemosensory proteins that are secreted in the lymph that bathes the dendrites of the SNs and are believed to solubilize and either transport the signaling molecules to membrane receptors or protect them from premature degradation (Vogt and Riddiford 1981; Pelosi et al. 2006). Although insect chemoreceptors and soluble chemosensory proteins are encoded by gene families exhibiting high gene turnover rates (see Sánchez-Gracia et al. 2011 for a comprehensive review), distant homologues of the members of these families have been identified in other arthropod lineages (Colbourne et al. 2011; Vieira and Rozas 2011; Chipman et al. 2014; Frías-López et al. 2015; Gulia-Nuss et al. 2016). Vertebrate functional counterparts of these gene families, however, are not evolutionarily related; indeed, the members of this subphylum use different molecules to perform the same general physiological function (Kaupp 2010).

Spiders comprise a highly diverse group of arthropods, including >45,000 described species (World Spider Catalog 2016), and are dominant predators in most terrestrial ecosystems. Given their potential as biological control agents as well as the engineering properties of silk and venom, these organisms are of great economic and medical relevance (Clarke et al. 2014). Because the Arachnida ancestors of these chelicerates colonized the land ~475 Ma, long after the split of the four major extant arthropod lineages (Rota-Stabelli et al. 2013), spiders are good models for comparative studies on the diverse strategies adopted by arthropod lineages during their independent adaptation to terrestrial environments. However, despite their biological and translational implications, there are relatively few genomic and transcriptomic studies conducted on these organisms compared with those conducted on insects, and studies on spiders almost exclusively focus on silk and venom research (Grbić et al. 2011; Clarke et al. 2014; Posnien et al. 2014; Sanggaard et al. 2014).

Spiders can detect volatile and nonvolatile compounds through specialized chemosensitive hairs distributed at the tips of various extremities and appendages, including legs and palps (Foelix 1970; Foelix and Chu-Wang 1973; Kronestedt 1979; Cerveira and Jackson 2012; Foelix et al. 2012). Nevertheless, the molecular nature of chelicerate chemoreceptors has remained elusive until recently. We and others have identified distant homologs of some insect gene families associated with chemosensation in the genomes of mites, ticks and spiders (Montagné et al. 2015; Gulia-Nuss et al. 2016), such as members of the gustatory (*Gr*) and ionotropic (*Ir*) receptor, and of the chemosensory protein (*Csp*), Niemann–Pick protein type C2 (*Npc2*) and sensory neuron membrane protein (*Snmp*) multigene families. In addition, chelicerates lack homologs of the typical insect olfactory receptor family *Ors*, which are thought to have originated later with the appearance of flying insects, and no *Obp* gene had been detected to date (Vieira and Rozas 2011; Chipman et al. 2014). Overall, available genomic studies suggest that the *Ir* gene family is responsible for smell not only in chelicerates but also in all nonneopteran arthropods (Croset et al. 2010; Colbourne et al. 2011; Chipman et al. 2014; Gulia-Nuss et al. 2016). Regarding taste, the presence of numerous copies of *Gr* and nonconserved *Ir* (a group of divergent IR proteins associated with gustatory function in insects, Croset et al. 2010) genes in chelicerate genomes clearly suggests that these families are responsible for contact chemoreception in this species.

Nevertheless, the simple comparative analysis of genomic sequences does not allow inferring which specific members of already known chemosensory families are involved in the different sensory modalities. Additionally, chelicerates could also use molecules completely different from those already known in insects during the water-to-land transition, which should also be different from those used by vertebrates (these molecules have also not been found in the available genome sequences); these uncharacterized genes (or annotated with incomplete gene models) would be not directly detectable only by comparative genomics. Instead, specific transcriptomic analyses of chemosensory tissues can provide useful insight into all these issues. Antennae-specific gene expression studies in lobsters and hermit crabs (Corey et al. 2013; Groh-Lunow et al. 2014), for example, have revealed the presence of several transcripts encoding IRs, supporting the active role in olfaction of this gene family in crustaceans. To gain insight into the specific proteins involved in chelicerate chemoreception, we recently performed a tissue-specific comparative transcriptomics study in the funnel-web spider *Macrothele calpeiana* (Frías-López et al. 2015). Unfortunately, we failed to detect the specific expression of *Ir* or *Gr* genes in the first pair of legs and in palps, the best candidate structures to hold olfactory hairs in chelicerates. This result might be caused by either the sedentary lifestyle of this mygalomorph spider, which may lead to a marginal role of chemical communication in this species, or the low sequencing coverage of this RNA-Seq study.

Here, in order to better characterize the chemosensory repertoire of a spider, we report a more comprehensive comparative transcriptomic analysis in an active nocturnal hunter spider, *Dysdera silvatica* Schmidt, 1981 (Araneae, Dysderidae) (fig. 1). This species, which is endemic to the Canary Islands, belongs to a genus characterized by long and protruding chelicerae used to capture and feed on woodlice (Crustacea: Isopoda: Oniscidea; fig. 1*B*). We have conducted a deep RNA-Seq experiment in four separated body parts, three of them likely containing chemosensitive hairs in spiders. Because the performance of the *de novo* assembly of short reads strongly depends on biological data (i.e., the complexity of the data is almost species specific), we first performed a comparative analysis among a set of commonly used software for transcriptome assembly. Based on the

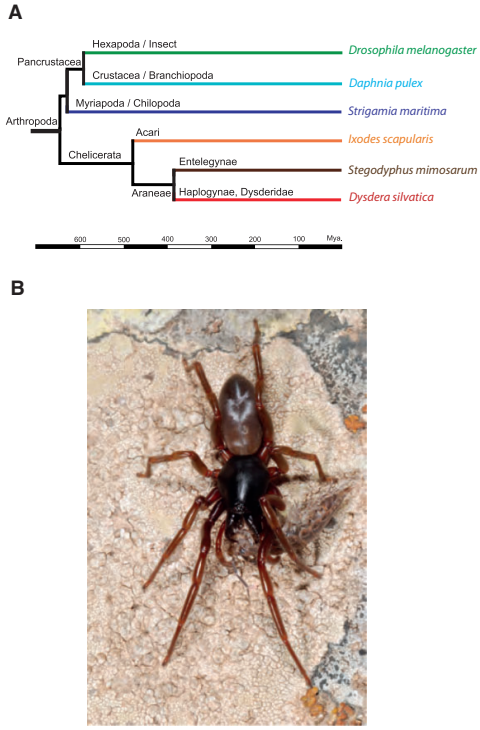


Fig. 1.—(*A*) Phylogenetic position of *Dysdera silvatica* within arthropods. Divergence times were obtained from TimeTree (Hedges et al. 2015). (*B*) *D. silvatica* feeding on a woodlouse.

best assembly and highly accurate functional annotations, we conducted a comparative analysis between the specific transcriptomes of the different body parts, emphasizing the detection of distinctive chemosensory profiles, especially in the palps and the first pair of legs, which has been reported to hold the peripheral olfactory structures in spiders. We then contextualized these results by applying a sound phylogenetic analysis including representative members of each arthropod chemosensory gene family.

We have identified several members of the *Ir* and *Gr* gene families specifically or differently expressed in some of the four surveyed transcriptomes (including a clear homolog of the co-receptor IR25a of *Drosophila melanogaster*) and some signs of chemosensory specialization across spider chemosensory structures. Moreover, we have also identified three genes distantly related to the insect *Obp* gene family and a new gene family encoding small secreted soluble proteins that might function as molecular carriers in the spider chemosensory system. We discuss these findings in the context of the origin and evolution of chemosensory gene families in arthropods and propose some candidate genes that may have an important chemoreceptor role in spiders.

Materials and Methods

Sample Collection, RNA Extraction and Library Preparation

We sequenced and analysed the transcriptome of four *D. silvatica* males (voucher specimens were deposited at the *Centre de Recursos de Biodiversitat Animal* of the Universitat de Barcelona under catalog numbers NMH2597-99 and NMH2601) collected from the Canary Islands, La Gomera and Las Tajoras (28.112736 N, 17.262511 W) in 2013. We used males because this sex has been shown to respond to sex-specific olfactory information (Nelson et al. 2012). We performed four separated RNA-Seq experiments, which included expressed sequences form the palps (*PALP*), the first pair of legs (*LEG#1*), all other pairs of legs (*LEG#234*)

and the remaining body structures (*REST*), henceforth referred to as experimental conditions. We dissected these body parts independently for each of the four males (after snap freezing in liquid nitrogen) and extracted the total RNA separately for each condition and sample using the RNeasy Mini kit (Qiagen, Venlo, The Netherlands) and TRIzol reagent (Invitrogen, Waltham, MA). We determined the amount and integrity of RNA using a Qubit Fluorometer (Life Technologies, Grand Island, NY) and Agilent 2100 Bioanalyzer (CCiTUB, Barcelona, Spain), respectively. We sequenced the transcriptome of each condition using the Illumina Genome Analyzer HiSeq 2000 (100 bp PE reads) according to the manufacturer's instructions (Illumina, San Diego, CA). Briefly, for each experimental condition, the mRNA was purified from 1 μg of total RNA using magnetic oligo(dT) beads and fragmented into small pieces. Double-stranded cDNA was synthesized with random hexamer (N6) primers (Illumina), and Illumina paired-end (PE) adapters were ligated to the ends of adenylated cDNA fragments. All library preparation steps and transcriptome sequencing were carried out in Macrogen Inc., Seoul, South Korea.

Raw Data Pre-Processing

Raw NGS data were pre-processed to eliminate all reads with a quality score ≤20 in at least the 30% of the read length and to remove reads with putative sequencing errors using NGSQCToolkit and SEECER v_0.1.3 (Patel and Jain 2012; Le et al. 2013). Before the assembly step, we performed an in silico normalization of filtered reads using Diginorm, an algorithm included in Trinity software (Haas et al. 2014). We set 50X as the targeted maximum coverage for the reads.

De Novo Transcriptome Assembly

First, to determine the best assembler for the *D. silvatica* RNA-Seq data, we compared the performance of five commonly used software programs in assembling the specific transcriptome of the experimental condition *REST*. We tested Trinity r2.1.1, Bridger r2014-12-01, SOAPdenovo-Trans release 1.03, Oases version_0.2.8, and ABySS version_1.3.7/trans-ABySS version1.4.8 (Birol et al. 2009; Schulz et al. 2012; Xie et al. 2014; Z. Chang et al. 2015). For this comparative analysis and depending on the specificities of the selected software (allowing single or multiple *k*-mer values), we applied several single *k*-mer lengths and *k*-mer ranges (see supplementary table S1, Supplementary Material online, for details).

After the assembly phase, we removed all contigs with evidence of contaminant sequences using the software Seqclean (ftp://occams.dfci.harvard.edu/pub/bio/tgi/software/; last accessed May 1, 2015) together with the sequences of the UniVec vector database and the genomes of *Escherichia coli, Pseudomonas aeruginosa, Staphylococcus aureus, Saccharomyces cerevisiae* and *Homo sapiens*. Clean contigs were then clustered into putative transcripts (analogous to the Trinity *components*). We determined the assembly performance of each software based on (1) the DETONATE score (Li et al. 2014), (2) the outcome of the assembled sequences in a set of sequence similarity and profile-based searches using different databases (see the "Results" section for more details), and (3) some commonly used descriptive statistics on assembly quality, namely the average sequence length, the N50, the maximum and minimum transcript lengths and the total bases in the assembly, calculated with the NGSQCToolkit software and some Perl scripts. All analyses were run in a 64-CPU machine with 750 Gb of RAM.

Protein Databases

We built two customized protein databases to assist the functional annotation of the *D. silvatica* transcriptome. The arthropodDB database contains the publicly available amino acid sequences of fully annotated proteins and protein models from a set of representative arthropod genomes and some appropriated external groups, along with their complete entry description, associated GO terms and InterPro identifiers (Ashburner et al. 2000; Mitchell et al. 2014). This database includes information for the following species: (1) the chelicerates *Ixodes scapularis* (Acari) *(Gulia-Nuss* et al. *2016)*, *Metaseiulus occidentalis* (Acari) (https://www.hgsc.bcm.edu/arthropods/western-orchard-predatory-mite-genome-project; last accessed May 1, 2015), *Tetranychus urticae* (Acari) (Grbić et al. 2011), *Mesobuthus martensii* (Scorpiones) *(Cao* et al. *2013)*, *Acanthoscurria geniculata* (Araneae, Theraphosidae) *(*Sanggaard et al. 2014*)*, *Stegodyphus mimosarum* (Araneae, Eresidae) *(*Sanggaard et al. 2014*)*, *Latrodectus hesperus* (Araneae) (https://www.hgsc.bcm.edu/arthropods/western-black-widow-spider-genome-project; last accessed May 1, 2015), *Loxosceles reclusa* (Araneae, Sicariidae) (https://www.hgsc.bcm.edu/arthropods/brown-recluse-spider-genome-project; last accessed May 1, 2015) and *Parasteatoda tepidariorum* (Araneae, Theridiidae) (https://www.hgsc.bcm.edu/arthropods/common-house-spider-genome-project; last accessed May 1, 2015); (2) the hexapods *D. melanogaster* (Diptera) (Adams et al. 2000), *Pediculus humanus* (Phthiraptera) *(*Kirkness et al. 2010*)* and *Bombyx mori* (Lepidoptera) *(*Mita et al. 2004*)*; (3) the crustacean *Daphnia pulex* (Branchiopoda) (Colbourne et al. 2011); (4) the myriapod *Strigamia maritima* (Chilopoda, Geophilomorpha) (Chipman et al. 2014); (5) the tardigrade *Hypsibius dujardini* (http://badger.bio.ed.ac.uk/H_dujardini; last accessed May 1, 2015); and (6) the nematode *Caenorhabditis elegans*. In the cases where there was no functional description or associated GO term (e.g., the protein models from *A. geniculata*, *L. hesperus*, *L. reclusa*, *M. martensii*, *M. occidentalis* and *P. tepidariorum*), we approximated the functional annotation using InterProScan version 5.4.47 (Jones et al. 2014).

The chemDB database contains the amino acid sequences and the functional information of all well-annotated members

of the *Or*, *Gr Ir*, *Csp*, *Obp*, *Npc2* and *Snmp* gene families from a representative set of insect species, namely *D. melanogaster*, *Tribolium castaneum* (Coleoptera), *Apis mellifera* (Hymenoptera) and *Acyrthosiphon pisum* (Hemiptera), and from the noninsect species included in arthropodDB. Moreover, we also included in chemDB some vertebrate odorant binding proteins and olfactory and taste receptors identified by the InterPro signatures IPR002448, IPR000725 and IPR007960, respectively (see supplementary table S1*B* in Frías-López et al. 2015). Furthermore, we progressively updated chemDB by adding to this database all novel members of these chemosensory families (the conceptual translation of the identified transcripts) characterized in *D. silvatica*.

Functional Annotation of the *D. silvatica* Transcripts

We applied a similarity-based search approach to assist the annotation of the *D. silvatica* transcriptome. We first used *BLASTx* to search the translated transcripts against the SwissProt and arthropodDB databases (BLAST v2.2.29; Altschul et al. 1990; Altschul 1997). To search against NCBI-nr, we used GHOSTZ version1.0.0; this software is much faster than *BLAST*, especially for large databases without a substantial reduction of sensibility (Suzuki et al. 2014). We improved the functional annotation by searching for the specific protein-domain signatures in translated transcriptome sequences using InterProScan (Jones et al. 2014). We predicted signal peptides and transmembrane helices with SignalP and TMHMM, respectively (Krogh et al. 2001; Petersen et al. 2011). To carry out the profile-based searches, we created custom HMM models, one for each chemosensory family included in chemDB. These models are based on multiple sequence alignments (MSA) built with the program *hmmalign* (HMMER 3.1b1 package; Eddy 2011) using the specific core Pfam profile as a guide.

We conducted a GO-enrichment analysis with the BLAST2GO term suite using all functionally annotated transcripts with an associated GO term (Conesa et al. 2005). Moreover, we also searched these functionally annotated transcripts for KEGG enzymes and pathways (Kanehisa and Goto 2000), for CEG (Core Eukaryotic Genes) (Parra et al. 2007; Parra et al. 2009) and for the list of housekeeping (HK) genes used in supplementary table S1*A* in Frías-López et al. (2015).

To characterize the chemosensory gene repertory of *D. silvatica*, we first used the proteins in chemDB as query sequences to search for putative homologs among spider transcripts (using *tBLASTn* search; *E*-value cutoff of 10^{-3}). We only considered as positives those hits covering at least 2/3 of the query sequence length or the 80% of the total subject sequence. Then, we conducted some additional searches based on our custom HMM models and the conceptual translation of *D. silvatica* transcripts as subject sequences (using *hmmer* and an i-*E*-value of 10^{-3}). The integration of the results from these different analyses provided us a highly curated and trustworthy set of *D. silvatica* chemosensory-related transcripts.

Expression Profiling across Experimental Conditions

The pre-processed reads of each experimental condition (*LEG#1*, *LEG#234*, *PALP*, and *REST*) were back aligned to the final reference transcriptome using Bowtie version 1.0.0 (Langmead et al. 2009). We used RSEM 1.2.19 software to obtain read counts and TMM-normalized FPKMs (i.e., trimmed mean of M values-normalized fragments per kb of exon per million reads mapped) per transcript (Li and Dewey 2011). For the analysis, we consider that a gene is actually expressed when the FPKM values are >0.01, a reasonable cut-off given the low expression levels reported for other arthropod chemoreceptor proteins (Zhang et al. 2014). For the differential expression analysis, we considered that our data represent a single biological replicate (Robinson et al. 2010) and used EdgeR version 3.6.8 to calculate the negative binomial dispersion across conditions from the read counts of HK genes (Robinson et al. 2010). The *P* values from the differential expression analysis were adjusted for the false discovery rate (FDR; Benjamini and Hochberg 1995).

Phylogenetic Analyses

The quality of the MSA is critical to obtain a reliable phylogenetic reconstruction. This issue is very problematic in the face of highly divergent sequences, as in our case. To minimize this problem, we applied a profile-guided MSA approach based on highly curated Pfam core profiles, which generated MSAs with better TCS scores than other MSA approaches (Chang et al. 2014; J.-M. Chang et al. 2015). We used RAxML version 8.2.1 and the WAG protein substitution model with rate heterogeneity among sites to determine the phylogenetic relationships among the members of each chemosensory gene family in arthropods (Whelan and Goldman 2001; Stamatakis 2014). Node support was estimated from 500 bootstrap replicates. All phylogenetic tree images were created using the iTOL webserver (Letunic and Bork 2007). Trees were rooted according to available phylogenetic information; otherwise, we applied a midpoint rooting.

Results

Evaluation of the Best *De Novo* Assembly for *D. silvatica* Data

We obtained 441.8 million reads across the four experimental conditions, which dropped to 418.2 million (94.7%) after removing low-quality reads (table 1). We used the 98.4 million reads of the *REST* condition to evaluate the best *de novo* transcriptome assembler for our specific data. We found that among the assemblers using a single *k-mer* value of 25, SOAPdenovo-Trans and Trinity produced the largest number of contigs and the lowest N50 values (supplementary table S1,

Table 1
Summary of RNA-Seq Data Assembly and Annotation

	PALP	*LEG#1*	*LEG#234*	*REST*	Total	Total aligned
Total raw reads	114,986,182	118,017,386	104,967,256	103,865,040	441,835,864	441,835,864
GC (%)	41.41	41.38	41.39	41.55	41.43	41.43
Total qualified reads	108,490,938	112,102,210	99,231,056	98,380,850	418,205,054	418,205,054
Transcripts	130,908	144,442	147,737	149,796	236,283	214,969
Unigene transcripts (UT)	93,283	104,004	106,966	109,335	170,846	154,427
UT average length (in bp)	1,027	956	943	932	702	751
UT maximum length (in bp)	26,709	26,709	26,709	26,709	26,709	26,709
HK UT	1,134	1,134	1,131	1,133	1,136	1,136
CEG UT (CEG genes)	766 (456)	766 (457)	775 (457)	759 (457)	807 (457)	804 (457)
UT with GO annotation	20,481	21,799	22,332	23,471	29,879	28,157
UT with Interpro domain	21,436	22,735	23,293	24,435	30,886	29,168
UT with KEGG annotation	3,313	3,409	3,444	3,599	3,895	3,817
UT with functional annotation[a]	21,567	22,874	23,438	24,600	31,091	29,359
UT with genomic annotation[b]	27,043	28,922	29,645	31,236	41,046	38,317

[a]GO, Interpro or KEGG annotation.
[b]GO, Interpro, KEGG annotation or BLAST hit.

Supplementary Material online). The assembly based on Bridger provided the second best RSEM-EVAL score (after Trinity) but produced contigs with more positive BLAST hits against CEG and SwissProt proteins with a 100% alignment length filtering with an *E*-value of 10^{-3}. Increasing the *k*-mer size had a disparate effect on the number of contigs and on the N50, but the resulting assemblies were generally worse than those generated using *k-mer* 25 (based on RSEM-EVAL scores and positive BLAST hits). Only the assemblies obtained in Bridger and Trinity with a *k*-mer of 31 outperformed their respective assemblies with a *k*-mer of 25. However, the multiple *k-mer* strategies implemented in Trans-Abyss and Oases yielded very different assembly qualities. Trans-Abyss produced a highly fragmented transcriptome (i.e., with a large number of very short contigs) that was clearly outperformed by Oases using the clustered option. Nevertheless, Oases performed worse than Bridger and Trinity (*k*-mer = 31) in terms of RSEM-EVAL scores and positive BLAST hits. Hence, although the Trinity assembly provided a lower RSEM-EVAL score, Bridger produced a very similar value of this parameter while performing better based on all other calculated statistics. Consequently, we selected Bridger with a *k*-mer of 31 as the best strategy for the *de novo* assembly of *D. silvatica* data and used the transcriptome from this software for further analyses.

The initial assembly from Bridger (using the reads from the four conditions) was formed by 236,283 contigs (after removing contaminant sequences), which decreased to 170,846 putative nonredundant transcripts after the clustering of isoforms (table 1). We identified 807 transcripts with significant BLAST hits against 457 out of the 458 CEGs, 454 of them with alignment lengths longer than the 60% of CEG target gene (234 with 100% of this length; supplementary table S2, Supplementary Material online). These results clearly demonstrate the completeness of the assembled transcriptome.

Functional Annotation of the *D. silvatica* Transcriptome

As expected, arthropodDB received the most significant positive BLAST hits with an *E*-value of 10^{-3} when using *D. silvatica* transcripts as queries (supplementary table S3, Supplementary Material online). Of these hits, 85% corresponded to chelicerate subjects; the spiders *A. geniculata* and *S. mimosarum* and the scorpion *M. martensii* were the most represented species (supplementary fig. S1, Supplementary Material online).

The most frequent GO terms associated with the *D. silvatica* transcripts were "metabolic" and "cellular processes" (biological process), as well as "binding" and "catalytic activities" (molecular function) (supplementary fig. S2, Supplementary Material online). Moreover, we found that 3,895 (out of the 29,879 transcripts with an associated GO term) showed significant positive BLAST hits against 136 different entries of the KEGG database (supplementary table S4, Supplementary Material online), with Purine metabolism (2,030 transcripts), Thiamine metabolism (1,053 transcripts) and Biosynthesis of antibiotics (454 transcripts including, e.g., some spider glutamate synthases and dehydrogenases) being the most represented pathways.

Condition-Specific Gene Expression Analysis

Our comparative analysis identified 57,282 transcripts expressed in all four conditions (37.1%) (fig. 2). The number of condition-specific transcripts in *LEG#1*, *PALP* and *LEG#234* was rather similar (7,446, 6,000 and 8,605, respectively) and was much higher in *REST* (14,414), which is easily explained by the much larger number of tissues and physiological functions included in this condition. In the absence of separated biological replicates, we used the expression profile of HK genes to estimate the approximate dispersion of mean

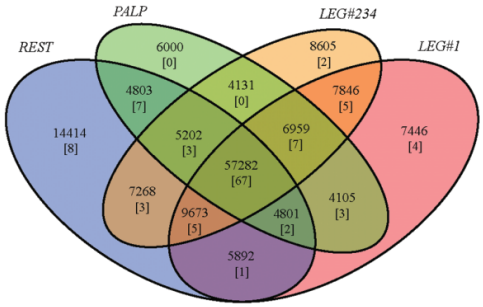


Fig. 2.—Venn diagram showing the total number of transcripts (154,427 transcripts) specifically expressed in each experimental condition and their intersections (red, orange, green and blue indicate *LEG#1*, *LEG#234*, *PALP* and *REST*, respectively). Numbers in brackets indicate putative chemosensory protein encoding transcripts (117 in total).

read counts across conditions to perform a rough differential expression analysis. The estimated dispersion across conditions of the 1,136 transcripts with significant positive BLAST hits to our set of HK genes (edgeR common dispersion value of 0.15) was used as the fold-change threshold for this analysis.

Our analyses show that *LEG#1* and *LEG#234* had rather similar transcriptomic profiles (supplementary fig. S3, Supplementary Material online). We found that only two transcripts were significantly overexpressed in *LEG#1* and the other two in *LEG#234*; taking these two conditions together, there were 27 overexpressed transcripts, none annotated as a chemosensory gene. These results contrast with those obtained in *PALP*, where 174 transcripts were significantly overexpressed. However, again, none of these transcripts encoded an annotated chemosensory function; they were enriched in signal peptide encoding sequences (Fisher's exact test, P value $= 2.63 \times 10^{-23}$), a feature characteristic of secreted proteins.

In addition, we found that the genes overexpressed in *PALP* were significantly enriched in the GO terms "metalloendopeptidase activity" (GO:0004222) and "proteolysis" (GO:0006508). In this specific tissue, these genes could be linked with the extra-oral digestion characteristic of these animals. However, we did not detect any GO term overrepresented in *LEG* or *REST*, and only 10 of the 27 genes significantly overexpressed in these structures had BLAST hits with an annotated sequence. Among these, we found genes encoding DNA-binding proteins, such as some transcription factors, hydrolases and proteins with transport activity.

Chemosensory Gene Families

To identify specific transcripts encoding chemosensory proteins in *D. silvatica*, we conducted additional exhaustive searches. We found many members of the *Gr*, *Ir*, *Npc2* and *Cd36-Snmp* families, as well as putative distant homologs of insect OBPs and one uncharacterized protein family that may be involved in chemosensory function in this spider. Nevertheless, we failed to find homologs of the *Csp* gene family, which is present in the genome of other chelicerates. As expected, the *D. silvatica* transcriptome did not encode insect OR proteins nor their vertebrate functional counterparts (supplementary table S5*A*, Supplementary Material online).

We identified 127 transcripts encoding IR/iGluR homologs (*Ir* transcripts), 57 exhibiting the specific domain signature of the ionotropic glutamate receptors (IPR001320). Some of these transcripts encoded some of the characteristic domains of the IR/iGluR proteins, such as the amino terminal (ATD-domain; PF01094), the ligand binding (LBD-domain; PF10613) and the ligand channel (LCD-domain; PF00060) (supplementary fig. S4, Supplementary Material online; see also Croset et al. 2010). Indeed, nine of them encoded all three domains, thus forming the typical complete iGluR structure, while 23 only had the two ligand-binding domains.

To understand the evolutionary diversification of the *Ir/iGluR* gene family in chelicerates, we carried out a protein domain-specific phylogenetic analysis. We used the information exclusively from the LCD domain because it is shared by all characterized arthropod IR/iGluR. For the analysis, we built an amino acid-based MSA including all *D. silvatica* transcript-coding LCD domains (70 transcripts) along with all reported sequences of this domain from *D. melanogaster*, *D. pulex*, *S. maritima*, *I. scapularis*, and *S. mimosarum* (i.e., in order to avoid large and unreadable trees, we included only one species per main arthropod lineage except for chelicerates, which were represented by a tick and a well annotated spider). We found that *D. silvatica* had representatives of all major IR/iGluR subfamilies, namely the AMPA, Kainate, NMDA (canonical iGluR subfamilies having all three Pfam domains), the two IR major subfamilies, the so called "conserved" IRs (encompassing the IR25a/IR8a members; having all three PFAM domains), and the remaining IR members (IR subfamily having only the LBD and LCD domains and that in *Drosophila* includes members with chemosensory function encompassing the so called "divergent" and the "antennal" IRs). In total, we identified 26 transcripts encoding canonical iGluR proteins plus another 44 encoding IRs (fig. 3 and supplementary fig. S5, Supplementary Material online), including a putative homolog of the highly conserved family of IR25a/IR8a proteins (transcript Dsil31989). Noticeably, this transcript is significantly overexpressed in *LEG#1* with respect to *REST* (~10 times more expression $-\log FC = 4$; $P < 0.01$ after FDR), although it also shows 2 and 4 times more FPKM values with respect to *PALP* and *LEG#234*, respectively (supplementary table S5*B*, Supplementary Material online).

Our phylogenetic analysis uncovered a set of *D. silvatica* transcripts phylogenetically related to some *D. melanogaster* antennal IRs, such as the IR21a (Dsil32714), the IR40a (Dsil150464) and the IR93a (Dsil55987, Dsil29850 and Dsil48134) proteins. These transcripts, however, did not show any clear differential expression pattern in *LEG#1* or

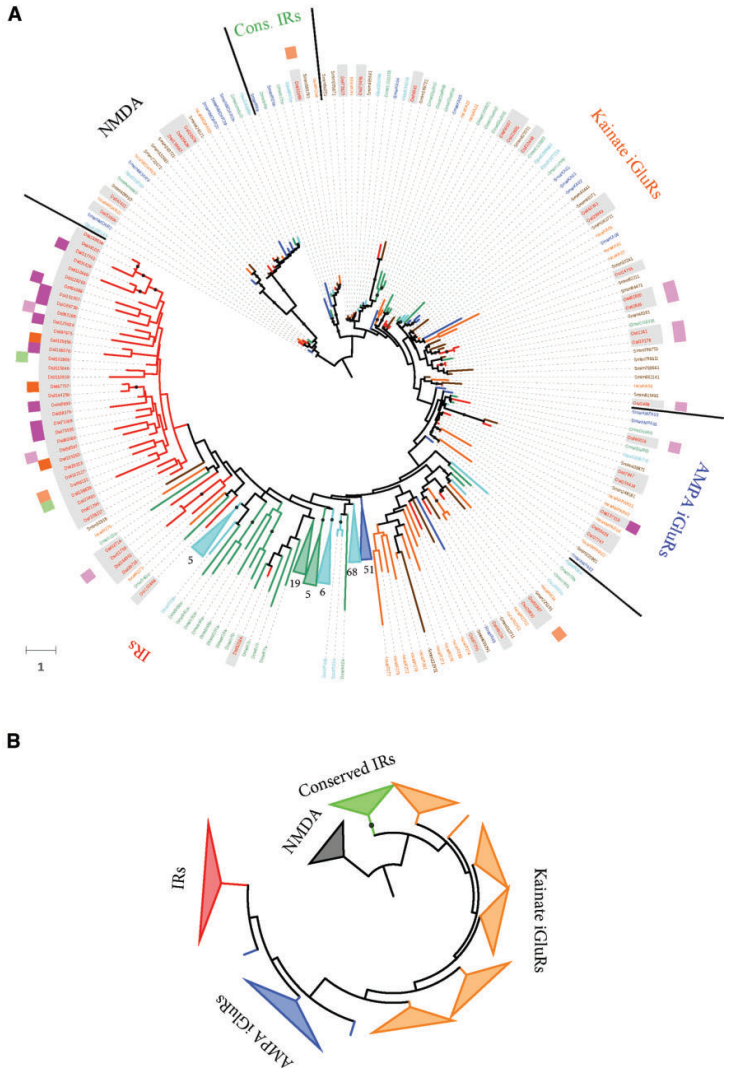


Fig. 3.—Maximum likelihood phylogenetic tree of the IR/iGluR proteins across arthropods. The tree is based on the MSA of the LCD domain (PF00060). (*A*) Sequences of *Drosophila melanogaster*, *Daphnia pulex*, *Strigamia maritima*, *Ixodes scapularis*, *Stegodyphus mimosarum* and *Dysdera silvatica* are depicted in green, light blue, dark blue, orange, brown and red, respectively. Additionally, the translation of the *D. silvatica* transcripts are shadowed in grey. Nodes with bootstrap support values >75% are shown as solid circles. Nodes with five or more sequences from the same species were collapsed; the actual number of collapsed branches is indicated in each case. The two surrounding circles provide information regarding the expression pattern of some *D. silvatica* genes. The most external circle indicates the genes specifically expressed in palps (PALP; in green), legs (both *LEG#1* and *LEG#234;* in pink) and palps and legs (*PALP*, *LEG#1* and *LEG#234;* in orange). The inner circle shows the genes overexpressed in these conditions using the same color codes but with two color intensities, one more intense color for overexpression levels >5× over *REST* and another lighter color for 2–5× overexpression values. The branch length scale is in numbers of amino acid substitutions per amino acid position. (*B*) Simplified phylogenetic tree highlighting the main *Ir* sub-families.

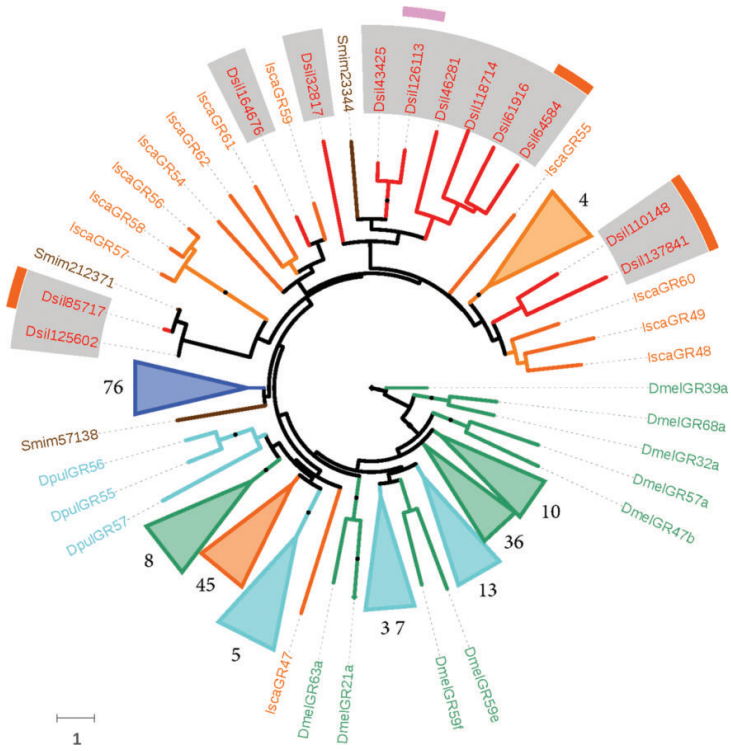


Fig. 4.—Maximum likelihood phylogenetic tree of the GR proteins across arthropods. Species names, node support features and surrounding circles are colored as in figure 3.

PALP, while two of them were clearly overexpressed in *REST*. Moreover, similarly to what occurs in other arthropods, many nonconserved IRs formed a species-specific monophyletic clade (33 transcripts). Interestingly, 11 of these receptors were condition specific, and 8 were overexpressed (or showed at least 2 times more FPKMs) in the examined appendages (i.e., *LEG#1*, *LEG#234* and *PALP* with respect to *REST*). Actually, *LEG#1* was the expression condition with the highest number of different nonconserved *Ir* transcripts; only 14 of the 43 nonconserved *Ir* members were not expressed in this appendage (supplementary table S5*B*, Supplementary Material online). Overall, the expression level of *Ir*s (including conserved *Ir*s) was lower than that of the *iGluR* transcripts.

We further identified 12 transcripts encoding GR proteins (*Gr* transcripts), although only four of them had one of the two specific InterPro signatures that characterize this family (7m_7, IPR013604 and Trehalose receptor, IPR009318). In addition, these 12 *Gr* transcripts were phylogenetically related to members of this family characterized in the spider *S. mimosarum* and in the deer tick *I. scapularis* (fig. 4 and supplementary fig. S6, Supplementary Material online). The expression levels of *D. silvatica Gr* genes were considerably low compared both with the overall expression levels and with the expression levels of other chemosensory families (supplementary table S5*C*, Supplementary Material online). Interestingly, only two *Gr* transcripts were condition specific (Dsil61916 and Dsil164676 in *REST*), and the other two were specifically expressed in both *LEG#1* and *PALP* (Dsil110148 and Dsil137841). The remaining *Gr* transcripts showed variable gene expression profiles across conditions, with some genes having a wide expression pattern and others being more restricted to particular conditions (supplementary table S5*C*, Supplementary Material online).

Our BLAST- and profile-based results revealed significant similarities between three spider transcripts and some insect

members of the *Obp* family (with *E*-values between 10^{-3} and 10^{-5}). The primary amino acid sequence and the cysteine pattern of the encoded proteins (hereafter designated OBP-like proteins) resembled those of OBPs and, one of them (Dsil553) showed a match to the PBP_GOBP InterPro domain (PBP_GOBP; IPR006170), uncovering a protein domain with folding features similar to those found in some insect OBPs. Using the three OBP-like sequences identified in the transcriptome of *D. silvatica* as a query in a BLASTp search against the NCBI-nr database (*E*-value of 10^{-3}), we detected six additional members of this novel family in the genomes of *S. mimosarum*, *I. scapularis* and *S. maritima* (two copies in each genome; fig. 5) but none in the annotated proteomes of crustaceans. The MSA of the nine copies identified in noninsect species and all characterized members of the *Obp* family in *D. melanogaster* and *Anopheles gambiae* would suggest that the *Obp*-like family is distantly related to the Minus-C *Obp* subfamily. Despite the particularly low sequence similarity and the large differences in protein length (not only between OBP-like and insect OBPs but also among OBP members), three different MSAs built with different alignment algorithms, i.e., MAFFT with the option L-INS-I (Katoh and Standley 2013), PROMAL3D (Pei et al. 2008) and PSI-coffee (Chang et al. 2012), yielded exactly the same pattern of cysteine homology in the region of the GOBP-PBP domain. Accordingly, with these MSAs, OBP-like proteins lacked the same two structurally relevant cysteines as insect Minus-C OBPs (except the *S. maritima* protein Smar010094 in the MAFFT alignment; supplementary fig. S7, Supplementary Material online). These results, however, must be taken with caution due to the fact that some OBP-like as well as several insect OBPs show large amino or carboxy terminal domains outside the conserved OBP domain, some of them including extra cysteines. If these cysteines are not correctly aligned in their true homologous positions, the interpretation of the cysteine pattern of OBP-like proteins could be erroneous.

We built a 3D protein model of both the conceptual translation of one of the *Obp*-like transcripts identified in *D. silvatica* (Dsil553) and of the *S. maritima* protein Smar010094 using the Phyre2 web portal (Kelley et al. 2015). As expected, the predicted models showed a globular structure very similar to that found in insect OBPs (fig. 6). In fact, the top 10 structural templates identified by the software and, therefore, the one selected for the final modeling (*A. gambiae* proteins OBP20 and OBP4 for Dsil553 and Smar010094, respectively) were insect OBPs. In addition, the models showed a high confidence in the region corresponding to the GOBP-PBP domain (56% and 59% of the query sequences were modeled with 89.2% and 81.6% confidence by the single highest scoring template, respectively). Remarkably, the amino acid alignment between Smar010094 and OBP4, used as a guide by Pyre2 for building the 3D model of this *S. maritima* OBP-like protein, coincided with the PROMAL3D and Psi-Coffee alignments but not with the MAFFT one (see above). Hence, we hypothesize that, given the wide expression of spider OBP-like across the four experimental conditions (supplementary table S5*D*, Supplementary Material online), these proteins, similar to those in insects, might be carriers of small soluble molecules acting in one or more physiological processes without ruling out a putative role in chemosensation.

We also identified 11 transcripts encoding putative NPC2 proteins, all of them having the characteristic IPR domain (MD-2-related lipid-recognition domain; IPR003172). The phylogenetic tree reconstructed from the MSA including these and other arthropod members of this family (including the members expressed in the antenna of *A. mellifera* and *Camponotus japonicus* (Ishida et al. 2014; Pelosi et al. 2014; fig. 7) uncovered a less dynamic gene family with neither large species-specific clades nor long branches. Nevertheless, internal node support was low and the precise phylogenetic relationships among arthropod NPC2s could not be determined with confidence. It is worth nothing, however, that this family underwent a moderate expansion in arthropods because it seems to be only one copy in both *C. elegans* and vertebrates. Only one putative *D. silvatica Npc2* transcript was *LEG#1* specific (Dsil113431), while two of them showed 11–4 times more FPKM in *PALP* (Dsil16636 and Dsil93094) and two others had 7 and 2 times more FPKM in *LEG#1* and *PALP* than in *REST* (Dsil56450 and Dsil793), respectively (supplementary table S5*E*, Supplementary Material online).

Finally, we identified 13 transcripts related to the *Cd36-Snmp* family, with 12 of them having the corresponding InterPro domain signature (CD36 antigen; IPR002159). Our phylogenetic analysis showed that *D. silvatica* had representatives of the three SNMP protein groups (Nichols and Vogt 2008; fig. 8), which would indicate that the origin of these subfamilies predated the diversification of the four major extant arthropod lineages. All four *D. silvatica Snmp* transcripts were similarly expressed in the four studied conditions, which would suggest either a nonchemosensory specific function of these proteins in spiders or a global general function within the chemosensory system (supplementary table S5*F*, Supplementary Material online).

A Novel Candidate Chemosensory Gene Family in Spiders

Furthermore, we conducted an exhaustive search on the 174 transcripts overexpressed in *LEG#1* and *PALP* to try to identify putative novel, previously uncharacterized spider olfactory chemosensory families. For this, we first searched for gene families (groups of 4 or more similar sequences) by performing a clustering analysis of the 174 transcripts with CD-HIT (Fu et al. 2012); then, we searched for the presence of a signal peptide or for signs of trans-membrane helices in the identified families. We found one family (with five copies) in which one member had the molecular hallmark of a signal peptide; the absence of such a mark in the other four members could be due to the failure to detect full-length transcripts in these

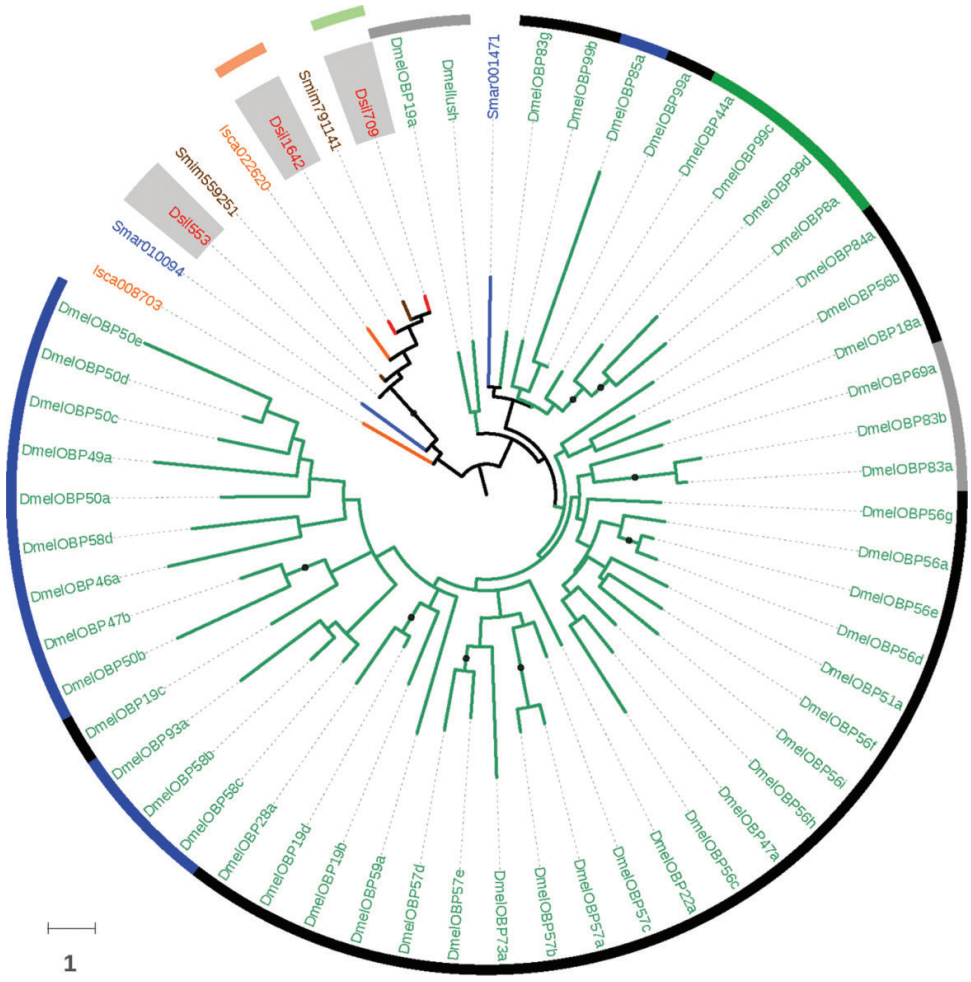


Fig. 5.—Maximum likelihood phylogenetic relationships of spider OBP-like and insect OBP proteins. Species names, node support features and surrounding circles are colored as in figure 3. The inner circle labels the previously defined OBP phylogenetic subfamilies (Classic, Minus-C, Plus-C and ABPII in black, green, blue and grey, respectively).

members (supplementary table S5*G*, Supplementary Material online). Using these five sequences as queries in a BLAST search against the complete *D. silvatica* transcriptome, we further detected seven more members of this family. New BLAST searches using all 12 proteins as queries identified homologous copies in other spiders but not in the genomes of either other chelicerate lineages or nonchelicerate species.

A preliminary phylogenetic analysis including all new identified sequences indicated that this family (supplementary fig. S8, Supplementary Material online) was highly dynamic, with several species-specific clades of CCPs (one of them including all *D. silvatica* copies) and no clear orthologous relationships across spiders. All these spider sequences, however, were annotated as uncharacterized proteins in these genomes.

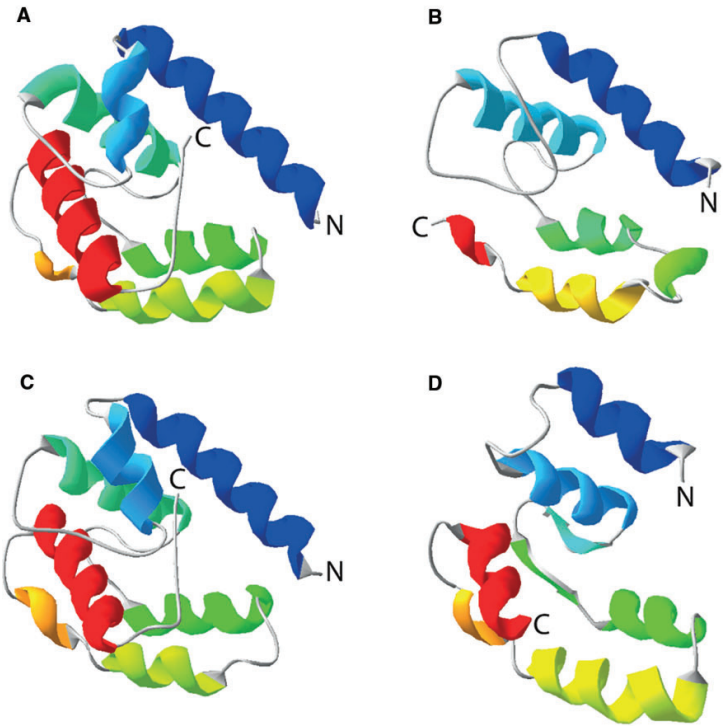


Fig. 6.—Predicted 3D structure of two OBP-like proteins. (*A*) Structure of *Anopheles gambiae* OBP20 (PDB 3V2L). (*B*) Structure of *A. gambiae* OBP4 (PDB 3Q8I). (*C*) 3D model of the protein encoded by the transcript Dsil553. (*D*) Predicted 3D model of the *Strigamia maritima* Smar010094 protein. PBD files were viewed and manipulated in Swiss-PdbViewer version 4.1 (Guex and Peitsch 1997).

The MSA of the members of this novel family revealed a conserved cysteine pattern similar to that observed in insect OBPs and CSPs. However, unlike the OBP-like proteins, we could not obtain a reliable 3D protein model of a member of this family in the Phyre2 webserver. The server was unable to identify reasonable templates with large alignment coverage for the modeling (all templates with confidences >15 had an alignment coverage <7%). We then used I-TASSER suite (Yang et al. 2015) to try to find template proteins of similar folds as our *D. silvatica* queries. Although two of the identified threading templates were OBPs, some artificially designed proteins were also included in the modeling, generating five highly heterogeneous folding models, most of them with unacceptable C-scores. Nevertheless, some of the estimated folding models showed a compact global structure that, along with the presence of a signal peptide and the gene expression data, would suggest that the members this novel gene family could also acts as carriers of small soluble molecules, as insect OBP do (hereinafter we will refer to this novel family as the *Ccp* gene family for candidate carrier protein family).

Discussion

A High-Quality *De Novo* Assembly of the *D. silvatica* Transcriptome

The key step to obtain a high-quality transcriptome is selecting the best *de novo* assembly strategy and software. Nevertheless, because most assemblers have been developed for specific NGS platforms or tested using reduced data sets with limited taxonomic coverage, it is very difficult to predict their performance with disparate datasets (Martin and Wang 2011). Obtaining a high-quality transcriptome depends on factors such as the organism (which determines DNA

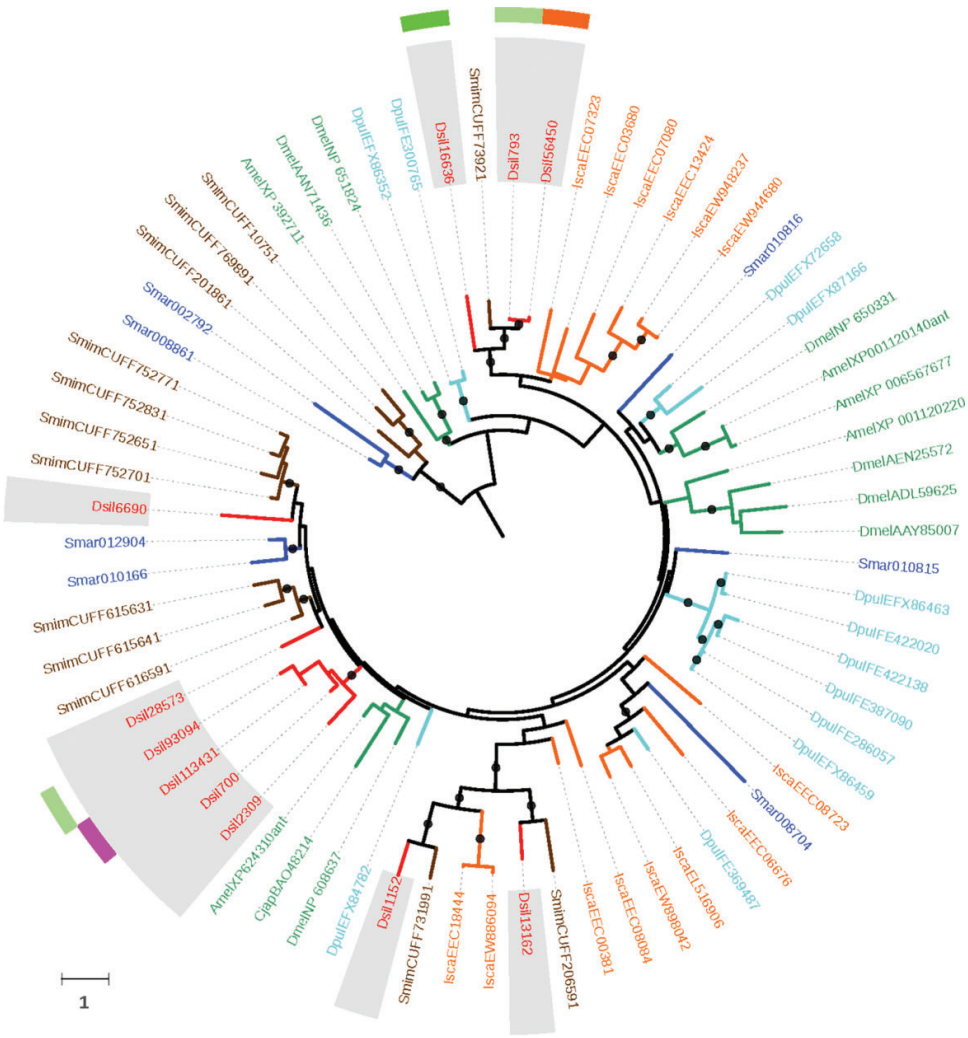


Fig. 7.—Maximum likelihood phylogenetic tree of the NPC2 proteins across arthropods. Species names, node support features and surrounding circles are colored as in figure 3. Sequences from *Apis mellifera* and *Camponotus japonicus* are colored in green.

complexity and heterozygosity levels), the read length and the sequencing depth. The best approach to determine the quality of different assemblies is to evaluate their accuracy (especially their completeness) in the context of a well-annotated, closely related reference genome (Marchant et al. 2015). Unfortunately, functionally annotated genomes of close relatives are usually not available for nonmodel organisms. In our case, the phylogenetically closest species with genome information, the spider *L. reclusa*, diverged from *D. silvatica* ~200 Ma (Binford et al. 2008), which prevented any reliable evaluation. To circumvent this limitation, we used a combination of two strategies to evaluate the performance of five competing assemblers, one based on information of the transcriptome completeness (using CEG and SwissProt databases as subjects)

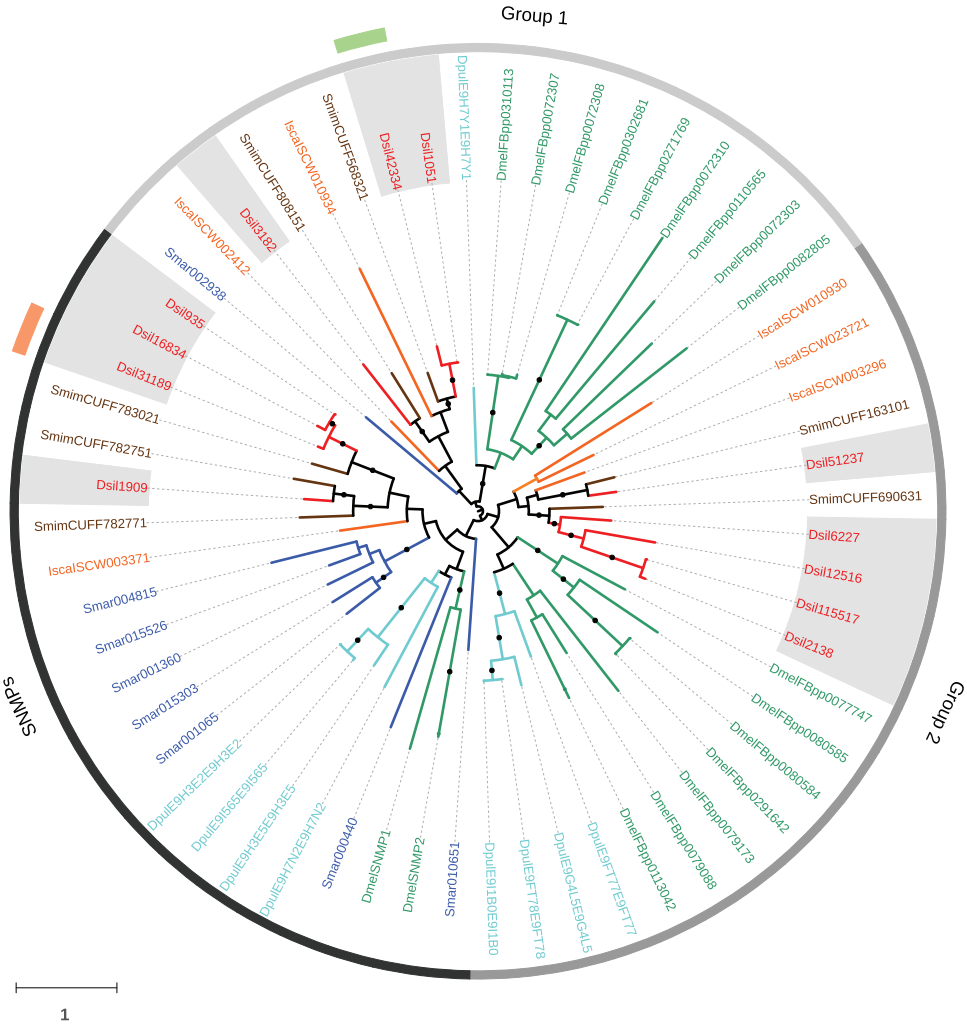


Fig. 8.—Maximum likelihood phylogenetic tree of CD36-SNMP proteins across arthropods. Species names, node support features and surrounding circles are colored as in figure 3. The inner circle shows the different subfamilies.

and the other based on some statistics measuring the assembly quality (Li et al. 2014). Using this combined strategy and after evaluating 11 assembly scenarios, we were able to obtain a high-quality assembly that probably covers most of the *D. silvatica* transcriptome and that has a large proportion of full-length transcripts.

A Comprehensive Annotated Transcriptome That Uncovers a Surprising Gene Loss in Chelicerata

The functional annotation of a *de novo* assembled transcriptome from a nonmodel organism is a daunting task, being usually slow and computationally intensive. The large number

of query sequences (transcripts) make similarity- and profile-based searches against general big databases, such as the NCBI-nr, very problematic, especially when using the free version of some software suites (e.g., BLAST2GO). Here, we used GHOSTZ instead of BLAST when searching against NCBI-nr, considerably reducing the computational time of the functional annotation step in >100 times, which is a relevant feature when testing assemblers in a comparative framework (i.e., a large number of independent annotations). Moreover, to increase the sensibility of the searches and reduce the computation time, we included only a representative set of phylogenetically close species to *D. silvatica* to build our specific databases (some annotated proteins are not yet available in NCBI-nr). Finally, we largely reduced the running time of the InterProScan searches (~10 times) by using only the Pfam database (Finn et al. 2014) as a query without a substantial loss in the number of positive hits.

Despite the exhaustive annotation process, a high number of *D. silvatica* transcripts (81.8%) could not be functionally annotated. These percentages, however, are commonly obtained in RNA-Seq studies and can be attributable to different causes. First, nonannotated transcripts are significantly shorter than annotated ones (P value = 2.2×10^{-16}), suggesting that many nonannotated transcripts are actually assembly errors or small fragments lacking any detectable protein domain signature (supplementary table S6, Supplementary Material online). Second, a fraction of these unannotated sequences could correspond to noncoding RNAs. Finally, the modest annotation of the genome of *L. reclusa*, the closest available relative to *D. silvatica*, could considerably reduce the success of our searches. In fact, an important number of *D. silvatica* transcripts without functional annotation (9,955 sequences) encoded proteins tagged as uncharacterized in the genome of *L. reclusa*.

A relevant result of our functional annotation of the *D. silvatica* transcriptome is the absence of a transcript encoding a Trehalase (KOG0602), the only gene of the CEG database not identified in the *D. silvatica* transcriptome. This gene seems to also be absent in the genomes of other chelicerates because we failed to detect it even using powerful profile-based approaches. Intriguingly, this protein is essential for insects (Shukla et al. 2015) not only because of its function as hydrolase but also for its involvement in the development of the optic lobe (Chen et al. 2014). Given that this gene is certainly present in the genome of all other major arthropod lineages as well as in the tardigrade *H. dujardini* and the nematode *C. elegans*, the most likely explanation for its absence is specific gene loss in the ancestor of chelicerates. The apparent absence of this gene in this lineage is interesting and clearly demands further investigation. The study of this gene loss, jointly with that of the set of uncharacterized proteins found in the *D. silvatica* transcriptome, will provide new insight into some important biological processes specific to chelicerates.

The Chemosensory Transcriptome of *D. silvatica*

Unlike our previous survey in the mygalomorph species *M. calpeiana* (Frías-López et al. 2015), here we identified several transcripts encoding members of chemosensory gene families in the four studied body parts, albeit with low expression levels. The different levels of success of the two studies could be related to the much higher sequencing depth (i.e., >10 Gbp sequenced per condition) of the *D. silvatica* RNA-Seq experiment.

As expected from the genome annotations of some chelicerate species, the transcriptome of *D. silvatica* did not contain genes related to the vertebrate chemoreceptors or odorant-binding protein families, ruling out the possibility that these or other similar families play any role in spider chemosensation. Similarly, we failed to detect members of the insect *Or* gene family, adding further evidence of the complete absence of this family in all arthropod lineages other than winged insects (Missbach et al. 2015). Moreover, despite the presence of members of the *Csp* gene family in some chelicerates and myriapods (Chipman et al. 2014; Qu et al. 2015; Gulia-Nuss et al. 2016), we did not identify any transcript encoding a protein with significant similarity to this family in *D. silvatica*. Although this negative result might be explained by sequencing or assembly limitations, *Csp* genes are also absent in all other spider genomes available in public repositories. We postulate that this gene family could have been lost early in the diversification of arachnids.

Candidate Spider Chemoreceptor Gene Families

Here, we identified a maximum of 12 transcripts encoding GR proteins (i.e., some of them may form part of the same gene), a number that may seem surprisingly small in comparison with the large number of *Gr* genes identified in the tick *I. scapularis* (62), the myriapod *S. maritima* (77) and the water flea *D. pulex* (58) genomes, for example. Nevertheless, given the underrepresentation of the chemosensitive hairs with respect to the total amount of tissue examined in each specific transcriptome, the identification and comprehensive annotation of the complete set of *Gr* genes are quite challenging in standard RNA-Seq studies (Zhang et al. 2014). In addition, some *Gr* genes do not necessarily have to be expressed at the precise moment (i.e., developmental stage or environmental condition) of the experiment (this can also be applied to all other chemosensory families). Therefore, the *D. silvatica* genome likely encodes many more members of this family, and the 12 transcripts found in this study are only a first preliminary subset of the gustatory repertoire of this spider. These molecules seem to be expressed across different spider body parts and some show specific expression in particular appendages, with groups of copies broadly expressed, other groups that are never found in particular appendages and others that show an opposite pattern of specificity. This combinatorial manner of expression is similar to that the described for the

*Gr*s in *Drosophila*, which would suggest analogous gustatory coding mechanisms in these two arthropods (Depetris-Chauvin et al. 2015; Joseph and Carlson 2015). The two phylogenetically related *Gr* genes specifically expressed in *LEG#1* and *PALP* (Dsil110148 and Dsil137841) could be involved in the detection of some ecologically relevant signals, for example, partial pressure of CO_2, in a similar way as some insect *Gr* specifically expressed in *D. melanogaster* antenna, although the proteins encoded by spider and insect transcripts are phylogenetically unrelated. In fact, all *Gr* transcripts detected in the *D. silvatica* transcriptome (including *LEG#1* and *PALP* specific sequences) are members of a monophyletic group of chelicerate receptors for which we have no functional information. However, some *Gr* transcripts are also overexpressed or even exclusively expressed in the transcriptome of *REST*. The encoded proteins might participate in other, nonchemosensory physiological functions, as has also been observed in insects (Joseph and Carlson 2015). Even so, we cannot rule out that they actually act as chemoreceptors in other body structures, apart from palps and legs, such as in the mouthparts, which are included in *REST* transcriptome.

Unlike *Gr*s, we have detected in *D. silvatica* a substantial number of sequences (127) encoding putative *Ir* transcripts, including a putative homolog of the conserved *Ir* subfamily *Ir25a/Ir8a* (Dsil31989). The phylogenetic analysis of the members of this family in arthropods clearly reflects the effect of the long-term birth-and-death process acting on most members of this family. Remarkably, this effect is almost unnoticeable in iGluR and in conserved IRs proteins, ratifying the marked differences in gene turnover rates between subfamilies. This highly dynamic evolution of nonconserved IR jointly with that reported for other proteins associated with contact chemoreception has been suggested as a proof of the high adaptive potential of the molecular components of the gustatory system in arthropods (see Torres-Oliva et al. 2016, and references therein). Interestingly, some of the 10 nonconserved IRs not included in the *D. silvatica*-specific clade are phylogenetically related to some *D. melanogaster* antennal IRs, including one member that presumably plays an important role in thermosensation (IR21a). Nevertheless, the expression profiles of these five transcripts do not provide clues regarding their possible role in spider chemosensation (i.e., they do not show any specific gene expression pattern across conditions). Although the putative spider homolog of the *Ir25a/Ir8a* subfamily is also expressed in all four conditions, it is much more abundant in *PALP*, *LEG#1* and *LEG#234*, and even significantly overexpressed in *LEG#1* with respect to *REST*. The IR25a and IR28a proteins are widely expressed in *Drosophila* olfactory sensilla (and in olfactory organs of other arthropods; Croset et al. 2010) and have been involved in the trafficking to the membrane of the other IR and in a co-receptor function of food-derived chemicals and humidity and temperature preferences. Thus, our results indicate that the first pair of legs of spiders could be relevant for the detection of amines and/or aldehydes as well as for determining favorable ranges of certain environmental variables (Silbering et al. 2011; Min et al. 2013; Enjin et al. 2016). Finally, and similar to that observed in for *Gr* transcripts, some members of the nonconserved *Ir* subfamily are also detected in *REST*, further supporting their involvement in other nonchemosensory functions or, alternatively, the presence of chemosensory structures in body parts other than legs or palps.

Evolution of the IR Family in Arthropods

Since our phylogenetic analysis includes highly diverged sequences, we applied for first time domain-specific HMM profiles to guide the MSA of chemosensory families. This strategy has been especially useful for the *Ir/iGluR* families, exploiting the evolutionary information of the conserved ligand channel domain (LCD domain) clearly shared by all known members. The inferred tree mirrors the same focal phylogenetic groups obtained in previous works (Croset et al. 2010). Most tree reconstructions show that (1) the Kainate and AMPA proteins are closely related, and AMPA likely a derived linage, (2) the subfamily of the conserved IRs is the sister group of these Kainate/AMPA receptors, and (3) NMDA sequences represent the first offshoot. However, there are some important differences between the present study and findings regarding the putative origin of the nonconserved IRs. This group of IRs, which forms a supported monophyletic group in all tree reconstructions, is more closely related to non-NMDA receptors than to the remaining iGluRs in our tree, which could indicate that they originated from a Kainate- or AMPA-like receptor. Nevertheless, the poor support of some internal nodes, probably due to alignment artifacts caused by the diverse domain structure of *Ir/iGluR* families, precludes making definitive conclusions about the origin of these highly divergent receptors.

Novel Classes of Candidate Transport Proteins in Chelicerates

Pelosi et al. (Pelosi et al. 2014) proposed that some members of the *Npc2* family might be involved in the transport and solubilization of semiochemicals in noninsect arthropods, constituting an alternative to the insect OBP and CSP proteins involved in the peripheral events of olfaction. Here, we show that the spider *D. silvatica* has a similar repertoire of *Npc2* genes to that found in other surveyed arthropods, which seems to be expanded in arachnids. We identified one member of this family specifically expressed in *LEG#1* that may be a good candidate to participate in odor detection in spiders; this transcript, however, showed a relatively low expression level, in contrast to the very high expression levels observed in insect *Obp* and *Csp* genes. Although the remaining members of the *Npc2* family might also have other chemoreceptor functions in *Dysdera*, most of them probably perform other important physiological functions, such as

cholesterol lipid binding and transport, which is the known function of these proteins in vertebrates (Storch and Xu 2009).

One unexpected and remarkable result is the expression in *D. silvatica* of at least three genes encoding proteins with a secondary structure, conserved cysteine pattern (revealed in the MSAs that include insect OBPs and characteristic of the Minus-C subfamily) and predicted folding similar to that of insect OBPs. In fact, our searches using these newly identified OBP-like proteins as a query revealed that chelicerates and myriapods, but not crustacean or insects, have some copies of this family. In the absence of confirmation by functional experiments and structural data, these results suggest that the *Obp* superfamily was already present in the arthropod ancestor. We cannot confirm whether putative ancestors were actually members of the Minus-C subfamily because this group of proteins is polyphyletic in the OBP tree (Vieira and Rozas 2011). Nevertheless, the fact that chelicerate and myriapod genomes only carry Minus-C *Obp* genes supports them as the ancestral arthropod *Obp*. In *D. melanogaster*, the Minus-C *Obp*s are highly expressed in several tissues other than the head, including adult carcass, testis, male accessory glands, spermatheca and some larval tissues (data from FlyAtlas project; Chintapalli et al. 2007). The wide expression levels of OBP-like genes across all four experimental conditions, together with their low gene turnover rates in chelicerates, also indicate essential and multiple functional roles of these putative small soluble carriers, regardless of their possible function in the chemosensory system.

Lastly, the newly identified *Ccp* family encodes a protein with a clear signal peptide that shows similar folding characteristics to those of insect OBPs. Interestingly, half of their members are overexpressed in the proposed spider olfactory organs. In this case, however, we only detected homologous copies in the genomes of arachnids, where the products are annotated as uncharacterized proteins. Thus, both the NPC2 copy and the proteins encoded by the *Ccp* family are good candidate chelicerate counterparts of the insect OBP and the CSP proteins, and their specific function clearly deserves further exploration.

In this study, we report the first comprehensive comparative transcriptomic analysis across different body structures of a spider, including those that most likely carry the chemosensory hairs. Our results indicate that, as in other noninsect arthropods, gustatory and ionotropic receptor families are the best candidate peripheral chemoreceptors in chelicerates. Additionally, we found some noteworthy differences in the specific pattern of gene expression of the members of these chemosensory families across different body structures, some of them involving the putative olfactory system-containing organs, which can indicate some specialization of chemosensory structures across the body of *D. silvatica*. In addition, we identified a protein family in chelicerates that seems to be distantly related to the insect *Obp* family and have characterized a new gene family of small secreted soluble proteins analogous to the insect OBPs or CSPs that could act as molecular carriers in this species. Finally, we provide the first complete and functionally annotated transcriptome of a polyphagous predator species of the genus *Dysdera*, which will provide valuable information for further studies on this group, and a list of candidate genes suitable for further functional dissection. Our results will help better establish the specific role and sensory modality of each of these new identified genes and gene families in spiders while providing new insight into the origin and evolution of the molecular components of the chemosensory system in arthropods.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by the Ministerio de Economía y Competitividad of Spain (BFU2010-15484, CGL2012-36863 and CGL2013-45211) and the Comissió Interdepartamental de Recerca I Innovació Tecnològica of Catalonia, Spain (2009SGR-1287, 2014SGR-1055 and 2014SGR1604). J.V. was supported by an FPI grant (Ministerio de Economía y Competitividad of Spain, BES-2014-068437), C.F.-L. by an IRBio fellowship, and A.S.-G. by a Beatriu de Pinós grant (Generalitat de Catalunya, 2010-BP-B 00175), and J.R. and M.A.A. were partially supported by ICREA Academia (Generalitat de Catalunya). We acknowledge the Garajonay National Parks for granting collection permits and helping with lodging and logistics during fieldwork.

Literature Cited

Adams MD, et al. 2000. The genome sequence of *Drosophila melanogaster*. Science 287:2185–2195.

Altschul S. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Ashburner M, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 25:25–29.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc. 57:289–300.

Binford GJ, et al. 2008. Phylogenetic relationships of *Loxosceles* and *Sicarius* spiders are consistent with Western Gondwanan vicariance. Mol Phylogenet Evol. 49:538–553.

Birol I, et al. 2009. De novo transcriptome assembly with ABySS. Bioinformatics 25:2872–2877.

Cao Z, et al. 2013. The genome of *Mesobuthus martensii* reveals a unique adaptation model of arthropods. Nat Commun. 4:2602.

Cerveira AM, Jackson RR. 2012. Love is in the air: olfaction-based mate-odour identification by jumping spiders from the genus *Cyrba*. J Ethol. 31:29–34.

Chang J-M, Di Tommaso P, Lefort V, Gascuel O, Notredame C. 2015. TCS: a web server for multiple sequence alignment evaluation and phylogenetic reconstruction. Nucleic Acids Res. 43:W3–W6.
Chang J-M, Di Tommaso P, Notredame C. 2014. TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. Mol Biol Evol. 31:1625–1637.
Chang JM, Di Tommaso P, Taly JF, Notredame C. 2012. Accurate multiple sequence alignment of transmembrane proteins with PSI-Coffee. BMC Bioinformatics 13(Suppl 4):S1.
Chang Z, et al. 2015. Bridger: a new framework for *de novo* transcriptome assembly using RNA-seq data. Genome Biol. 16:1–10.
Chen EA, et al. 2014. Effect of RNA integrity on uniquely mapped reads in RNA-Seq. BMC Res Notes 7:753.
Chintapalli VR, Wang J, Dow JAT. 2007. Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. Nat Genet. 39:715–720.
Chipman AD, et al. 2014. The first myriapod genome sequence reveals conservative arthropod gene content and genome organisation in the centipede *Strigamia maritima*. PLoS Biol. 12:e1002005.
Clarke TH, et al. 2014. Multi-tissue transcriptomics of the black widow spider reveals expansions, co-options, and functional processes of the silk gland gene toolkit. BMC Genomics 15:365.
Colbourne JK, et al. 2011. The ecoresponsive genome of *Daphnia pulex*. Science 331:555–561.
Conesa A, et al. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674–3676.
Corey EA, Bobkov Y, Ukhanov K, Ache BW. 2013. Ionotropic crustacean olfactory receptors. PLoS One 8:e60551.
Croset V, et al. 2010. Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. PLoS Genet. 6:e1001064.
Depetris-Chauvin A, Galagovsky D, Grosjean Y. 2015. Chemicals and chemoreceptors: ecologically relevant signals driving behavior in *Drosophila*. Front Ecol Evol. 3:41.
Eddy SR. 2011. Accelerated profile HMM searches. PLoS Comput. Biol. 7:e1002195.
Enjin A, et al. 2016. Humidity sensing in *Drosophila*. Curr Biol. 26:1352–1358.
Finn RD, et al. 2014. Pfam: the protein families database. Nucleic Acids Res. 42:D222–D230.
Foelix RF, Chu-Wang IW. 1973. The morphology of spider sensilla. II. Chemoreceptors. Tissue Cell 5:461–478.
Foelix RF, Rast B, Peattie AM. 2012. Silk secretion from tarantula feet revisited: alleged spigots are probably chemoreceptors. J Exp Biol. 215:1084–1089.
Foelix RF. 1970. Chemosensitive hairs in spiders. J Morphol. 132:313–333.
Frías-López C, et al. 2015. Comparative analysis of tissue-specific transcriptomes in the funnel-web spider *Macrothele calpeiana* (Araneae, Hexathelidae). Peer J. 3:e1064.
Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152.
Grbić M, et al. 2011. The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. Nature 479:487–492.
Groh-Lunow KC, Getahun MN, Grosse-Wilde E, Hansson BS. 2014. Expression of ionotropic receptors in terrestrial hermit crab's olfactory sensory neurons. Front Cell Neurosci. 8:1–12.
Guex N, Peitsch MC. 1997. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 18:2714–2723.
Gulia-Nuss M, et al. 2016. Genomic insights into the *Ixodes scapularis* tick vector of Lyme disease. Nat Commun. 7:10507.
Haas BJ, et al. 2014. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat Protoc. 8:1–43.
Hedges SB, Marin J, Suleski M, Paymer M, Kumar S. 2015. Tree of life reveals clock-like speciation and diversification. Mol Biol Evol. 32:835–845.
Ishida Y, et al. 2014. Niemann-Pick type C2 protein mediating chemical communication in the worker ant. Proc Natl Acad Sci U S A. 111:3847–3852.
Jones P, et al. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240.
Joseph RM, Carlson JR. 2015. Drosophila chemoreceptors: a molecular interface between the chemical world and the brain. Trends Genet. 31:683–695.
Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28:27–30.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30:772–780.
Kaupp UB. 2010. Olfactory signalling in vertebrates and insects: differences and commonalities. Nat Rev Neurosci. 11:188.
Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 10:845–858.
Kirkness EF, et al. 2010. Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. Proc Natl Acad Sci U S A. 107:12168–12173.
Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 305:567–580.
Kronestedt T. 1979. Study on chemosensitive hairs in wolf spiders (Araneae, Lycosidae) by scanning electron microscopy. Zool Scr. 8:279–285.
Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10:R25.
Le H-S, Schulz MH, McCauley BM, Hinman VF, Bar-Joseph Z. 2013. Probabilistic error correction for RNA sequencing. Nucleic Acids Res. 41:e109.
Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23:127–128.
Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12:323.
Li B, et al. 2014. Evaluation of *de novo* transcriptome assemblies from RNA-Seq data. Genome Biol. 15:553.
Marchant A, et al. 2015. Comparing *de novo* and reference-based transcriptome assembly strategies by applying them to the blood-sucking bug *Rhodnius prolixus*. Insect Biochem Mol Biol. 69:25–33.
Martin J. a, Wang Z. 2011. Next-generation transcriptome assembly. Nat Rev Genet. 12:671–682.
Min S, Ai M, Shin SA, Suh GSB. 2013. Dedicated olfactory neurons mediating attraction behavior to ammonia and amines in *Drosophila*. Proc Natl Acad Sci U S A. 110:E1321–E1329.
Missbach C, Vogel H, Hansson BS, Groβe-Wilde E. 2015. Identification of odorant binding proteins and chemosensory proteins in antennal transcriptomes of the jumping bristletail *Lepismachilis y-signata* and the firebrat *Thermobia domestica*: evidence for an independent OBP-OR origin. Chem Senses 40:615–626.
Mita K, et al. 2004. The genome sequence of silkworm, *Bombyx mori*. DNA Res. 11:27–35.
Mitchell A, et al. 2014. The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res. 43: D213–D221.
Montagné N, de Fouchier A, Newcomb RD, Jacquin-Joly E. 2015. Advances in the identification and characterization of olfactory receptors in insects. Prog Mol Biol Transl Sci. 130:55–80.

Nelson XJ, Warui CM, Jackson RR. 2012. Widespread reliance on olfactory sex and species identification by lyssomanine and spartaeine jumping spiders. Biol J Linn Soc. 107:664–677.
Nichols Z, Vogt RG. 2008. The SNMP/CD36 gene family in Diptera, Hymenoptera and Coleoptera: *Drosophila melanogaster*, *D. pseudoobscura*, *Anopheles gambiae*, *Aedes aegypti*, *Apis mellifera*, and *Tribolium castaneum*. Insect Biochem Mol Biol. 38:398–415.
Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23:1061–1067.
Parra G, Bradnam K, Ning Z, Keane T, Korf I. 2009. Assessing the gene space in draft genomes. Nucleic Acids Res. 37:289–297.
Patel RK, Jain M. 2012. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One 7:e30619.
Pei J, Kim BH, Grishin VN. 2008. PROMALS3D: a tool for multiple sequence and structure alignment. Nucleic Acids Res. 36:2295–2300.
Pelosi P, Iovinella I, Felicioli A, Dani FR. 2014. Soluble proteins of chemical communication: an overview across arthropods. Front Physiol. 5:320.
Pelosi P, Zhou J-J, Ban LP, Calvello M. 2006. Soluble proteins in insect chemical communication. Cell Mol Life Sci. 63:1658–1676.
Pelosi P. 1996. Perireceptor events in olfaction. J Neurobiol. 30:3–19.
Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8:785–786.
Posnien N, et al. 2014. A comprehensive reference transcriptome resource for the common house spider *Parasteatoda tepidariorum*. PLoS One 9:e104885.
Qu S-X, Ma L, Li H-P, Song J-D, Hong X-Y. 2015. Chemosensory proteins involved in host recognition in the stored food mite *Tyrophagus putrescentiae*. Pest Manag Sci. 72(8):1508–1516.
Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140.
Rota-Stabelli O, Daley AC, Pisani D. 2013. Molecular timetrees reveal a Cambrian colonization of land and a new scenario for ecdysozoan evolution. Curr Biol. 23:392–398.
Sánchez-Gracia A, Vieira FG, Almeida FC, Rozas J. 2011. Comparative genomics of the major chemosensory gene families in arthropods. Encycl Life Sci. 3:476–490.
Sanggaard KW, et al. 2014. Spider genomes provide insight into composition and evolution of venom and silk. Nat Commun. 5:3765.
Schulz MH, Zerbino DR, Vingron M, Birney E. 2012. Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28:1086–1092.
Shanbhag SR, et al. 2001. Expression mosaic of odorant-binding proteins in *Drosophila* olfactory organs. Microsc Res Tech. 55:297–306.
Shukla E, Thorat LJ, Nath BB, Gaikwad SM. 2015. Insect trehalase: physiological significance and potential applications. Glycobiology 25:357–367.
Silbering AF, et al. 2011. Complementary function and integrated wiring of the evolutionarily distinct *Drosophila* olfactory subsystems. J Neurosci. 31:13357–13375.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.
Storch J, Xu Z. 2009. Niemann-Pick C2 (NPC2) and intracellular cholesterol trafficking. Biochim Biophys Acta. 1791:671–678.
Suzuki S, Kakuta M, Ishida T, Akiyama Y. 2014. Faster sequence homology searches by clustering subsequences. Bioinformatics 31:1183–1190.
Torres-Oliva M, Almeida FC, Sánchez-Gracia A, Rozas J. 2016. Comparative genomics uncovers unique gene turnover and evolutionary rates in a gene family involved in the detection of insect cuticular pheromones. Genome Biol Evol. 8:1734–1747.
Vieira FG, Rozas J. 2011. Comparative genomics of the odorant-binding and chemosensory protein gene families across the arthropoda: origin and evolutionary history of the chemosensory system. Genome Biol Evol. 3:476–490.
Vogt RG, Riddiford LM. 1981. Pheromone binding and inactivation by moth antennae. Nature 293:161–163.
Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 18:691–699.
Whiteman NK, Pierce NE. 2008. Delicious poison: genetics of *Drosophila* host plant preference. Trends Ecol Evol. 23:473–478.
World Spider Catalog. 2016. World Spider Catalog. Nat. Hist. Museum Bern:online at http://wsc.nmbe.ch; version 17.0.
Xie Y, et al. 2014. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics 30:1660–1666.
Yang J, et al. 2015. The I-TASSER suite: protein structure and function prediction. Nat Methods 12:7–8.
Zhang Y, Zheng Y, Li D, Fan Y. 2014. Transcriptomics and identification of the chemoreceptor superfamily of the pupal parasitoid of the oriental fruit fly, *Spalangia endius* Walker (Hymenoptera: Pteromalidae). PLoS One 9:e87800.

Associate editor: Davide Pisani

Evolution of chemosensory gene families in arthropods: Insight from the first inclusive comparative transcriptome analysis across spider appendages

Vizueta J., Frías-López C., Macías-Hernández N., Arnedo M.A., Sánchez-Gracia A. and Rozas J.

Supplementary Material

Table S1. Comparative of the different *de novo* assembler performance using the *REST* condition reads.

Statistics \ Assemblies	Bridger	Trinity r2.1.1	Oases	Oases > 200	SOAPdenovoTrans	SOAPdenovoTrans > 200
k-mer	25	25	25	25	25	25
RSEM-EVAL Score	-3,714,451,296	-3,598,445,172	-4,721,160,361	-4,422,449,636	-8,172,293,402	-8,061,800,367
Number of contigs	131,540	170,554	127,296	126,961	220,817	83,994
Total bases count	139,081,034	122,182,956	153,427,486	153,361,192	72,667,908	54,054,402
Min. sequence length	201	224	138	201	100	201
Max. sequence length	23,555	20,914	27,574	27,574	17,327	17,327
Average sequence length	1,057.33	716.39	1,205.28	1,207.94	329.09	643.55
N50	1,934	1,103	2,019	2,020	555	991
SwissProt blastx hits count	10,316	11,663	8,358	8,358	7,855	7,741
" with 100% coverage	3,202	2,872	3,007	3,007	1,515	1,515
" with at least 80% coverage	5,356	5,084	5,014	5,014	3,106	3,103
" with at least 50% coverage	7,362	7,547	6,723	6,723	5,384	5,372
CEG blastx hits count	456	456	449	449	448	444
" with 100% coverage	235	213	225	225	107	107
" with at least 80% coverage	434	411	418	418	267	267
" with at least 50% coverage	454	454	448	448	410	410

Statistics \ Assemblies	Bridger	Trinity r2.1.1	SOAPdenovoTrans	SOAPdenovoTrans > 200	Oases
k-mer	31	31	50	50	61
RSEM-EVAL Score	-3,208,396,650	-2,940,345,348	-8,806,853,428	-8,766,461,767	-5,038,928,888
Number of contigs	127,865	185,462	349,136	104,984	80,219
Total bases count	133,568,439	127,219,697	84,137,105	51,848,141	99,744,321
Min. sequence length	201	201	100	201	200
Max. sequence length	26,494	23,266	18,766	18,766	26,785
Average sequence length	1,044.49	685.96	240.99	493.87	1,243.00
N50	2,007	1,105	266	621	2,122
SwissProt blastx hits count	10,272	11,288	8,359	8,002	7,474
" with 100% coverage	3,224	3,051	809	809	2,188
" with at least 80% coverage	5,345	5,196	1,949	1,945	3,982
" with at least 50% coverage	7,343	7,421	4,340	4,317	5,669
CEG blastx hits count	456	456	440	426	443
" with 100% coverage	236	226	51	51	164
" with at least 80% coverage	432	428	164	164	370
" with at least 50% coverage	454	455	345	342	436

continued on next page

continued from previous page

Statistics \ Assemblies	Oases Multiple *k-mer*	Oases Multiple *k-mer* > *200*	Oases Multiple *k-mer* Clustered	Oases Multiple K-*mer* Clustered > 200	ABySS Multiple *k-mer*	ABySS Multiple *k-mer* > *200*
k-mer	19-63	19-63	19-63	19-63	19-63	19-63
RSEM-EVAL Score	-4,767,928,528	-4,444,624,482	-4,288,780,585	-3,974,812,871	-5,631,145,331	-5,393,431,050
Number of contigs	514,647	513,181	116,182	115,429	656,439	331,677
Total bases count	755,317,444	755,059,396	131,701,494	131,571,819	261,116,506	215,709,553
Min. sequence length	100	201	100	201	101	201
Max. sequence length	27,574	27,574	27,574	27,574	12,639	12,639
Average sequence length	1,467.64	1,471.33	1,133.58	1,139.85	397.78	650.36
N50	2,286	2,287	1,873	1,876	663	845
SwissProt blastx hits count	10,095	10,094	8,161	8,160	12,761	12,470
" with 100% coverage	3,295	3,295	2,943	2,943	1,919	1,919
" with at least 80% coverage	5,520	5,520	4,929	4,929	3,985	1,983
" with at least 50% coverage	7,741	7,741	6,575	6,575	7,660	7,636
CEG blastx hits count	454	454	454	454	451	449
" with 100% coverage	240	240	224	224	144	144
" with at least 80% coverage	433	433	429	429	307	307
" with at least 50% coverage	454	453	453	453	430	427

Table S2. Coverage distribution of CEG blastx hits

Percentage of coverage	Number of Hits	Number of accumulated hits
100	234	234
90	145	379
80	50	429
70	18	447
60	7	454
50	1	455
40	1	456
30	1	457
20	0	457
10	0	457
0	0	457

The complete CEG database includes 458 genes

Table S3. Hits of *D. silvatica* transcripts in the different databases

	Total	ArthropodaDB	NCBI-nr	SwissProt	All DB hits combined
Transcripts	170,846	39,646	30,860	18,417	41,046
Percentage		23.21%	18.06%	10.78%	24.03%

Table S4. KEGG pathways associated with *D. silvatica* transcripts. Available at *Genome Biology and Evolution* online https://doi.org/10.1093/gbe/evw296

Table S5A. Summary of the number of transcripts encoding chemosensory gene families in *D. silvatica*

Chemosensory Family	Transcripts	Domain Average Length (#AA)	Protein Average Length (#AA)	Genes the corresponding Interpro Domain
OR	0	0	0	0
IR	127	210	252	57
GR	12	172	183	4
CSP	0	0	0	0
OBPs-like	3	82	170	1
NPC2	11	137	166	11
CD36-SNMP	13	340	403	12
CCPs	12	-	107	-

#AA, number of amino acids

Table S5B. Expression levels of transcripts encoding chemosensory genes across conditions

Chemosensory Gene family	Subfamily	Transcript	Length	Location	FPKM			
					PALP	*LEG#1*	*LEG#234*	*REST*
IR/iGluR	Conserved IR	IR25a-Dsil31989	1,514	*PALP, LEG#1, LEG#234, REST*	0.418	0.933	0.289	0.093
IR/iGluR	IR	Dsil48134	1,081	*PALP, LEG#1, LEG#234, REST*	0.271	0.509	0.14	0.533
IR/iGluR	IR	Dsil29850	1,194	*PALP, LEG#1, LEG#234, REST*	1.194	0.416	0.405	1.217
IR/iGluR	IR	Dsil55987	999	*PALP, LEG#234, REST*	0.197	0	0.314	0.07
IR/iGluR	IR	Dsil87055	1,233	*REST*	0	0	0	0.661
IR/iGluR	IR	Dsil48102	1,130	*PALP, LEG#1, LEG#234, REST*	0.037	0.076	0.025	0.255
IR/iGluR	IR	Dsil117743	726	*LEG#1*	0	0.373	0	0
IR/iGluR	IR	Dsil35628	3,101	*PALP, LEG#1, LEG#234, REST*	0.234	0.271	0.297	0.417
IR/iGluR	IR	Dsil110440	564	*LEG#1, LEG#234, REST*	0	0.144	0.107	0.162
IR/iGluR	IR	Dsil126283	886	*PALP, LEG#1, LEG#234, REST*	0.049	0.127	0.033	0.081
IR/iGluR	IR	Dsil84368	746	*LEG#1, LEG#234, REST*	0	0.297	0.149	0.058
IR/iGluR	IR	Dsil151007	264	*LEG#234*	0	0	0.24	0
IR/iGluR	IR	Dsil166736	218	*LEG#234*	0	0	0.463	0
IR/iGluR	IR	Dsil83188	1,003	*PALP, LEG#1, LEG#234, REST*	0.111	0.246	0.174	0.104
IR/iGluR	IR	Dsil125924	391	*LEG#1, LEG#234*	0	0.178	0.099	0
IR/iGluR	IR	Dsil97975	607	*REST*	0	0	0	0.348
IR/iGluR	IR	Dsil135856	337	*PALP, LEG#1, LEG#234*	0.197	0.237	0.132	0
IR/iGluR	IR	Dsil106076	385	*LEG#1*	0	0.365	0	0
IR/iGluR	IR	Dsil101866	594	*PALP, LEG#234, REST*	0.295	0	0.099	0.07
IR/iGluR	IR	Dsil115648	383	*LEG#234, REST*	0	0	0.099	0.301
IR/iGluR	IR	Dsil110939	632	*PALP, LEG#1, REST*	0.074	0.246	0	0.197
IR/iGluR	IR	Dsil67707	822	*PALP, LEG#1*	0.098	0.119	0	0
IR/iGluR	IR	Dsil144256	264	*PALP, REST*	0.357	0	0	0.707
IR/iGluR	IR	Dsil40990	1,584	*LEG#1, LEG#234*	0	0.314	0.107	0
IR/iGluR	IR	Dsil38374	880	*LEG#1, LEG#234, REST*	0	0.051	0.033	0.128
IR/iGluR	IR	Dsil71008	976	*LEG#1, LEG#234*	0	0.416	0.132	0
IR/iGluR	IR	Dsil73595	279	*LEG#1*	0	0.365	0	0
IR/iGluR	IR	Dsil80380	721	*PALP, LEG#1, LEG#234, REST*	0.062	0.102	0.074	0.058
IR/iGluR	IR	Dsil58597	1,202	*PALP, LEG#1, LEG#234, REST*	0.086	0.288	0.124	0.093
IR/iGluR	IR	Dsil105263	1,341	*PALP, LEG#1, LEG#234*	0.111	0.076	0.091	0
IR/iGluR	IR	Dsil29323	1,113	*PALP, LEG#1, LEG#234, REST*	0.037	0.22	0.066	0.197
IR/iGluR	IR	Dsil102127	231	*LEG#234, REST*	0	0	0.372	0.533
IR/iGluR	IR	Dsil46131	2,597	*PALP, LEG#1, LEG#234, REST*	0.049	0.305	0.364	0.023
IR/iGluR	IR	Dsil138839	532	*PALP, LEG#234, REST*	0.258	0	0.058	0.081
IR/iGluR	IR	Dsil23465	382	*PALP, REST*	0.148	0	0	1.507
IR/iGluR	IR	Dsil21299	1,945	*PALP, LEG#1, LEG#234, REST*	0.234	0.059	0.14	3.258
IR/iGluR	IR	Dsil158217	221	*REST*	0	0	0	1.264
IR/iGluR	IR	Dsil32714	3,334	*PALP, LEG#1, LEG#234, REST*	0.148	0.051	0.017	1.125
IR/iGluR	IR	Dsil31759	2,478	*PALP, LEG#1, LEG#234, REST*	0.394	0.585	0.595	0.22
IR/iGluR	IR	Dsil144061	313	*PALP, REST*	0.455	0	0	0.232
IR/iGluR	IR	Dsil150464	242	*REST*	0	0	0	0.916
IR/iGluR	IR	Dsil92084	496	*REST, LEG#1*	0	0.059	0	0.29
IR/iGluR	IR	Dsil169934	201	*PALP, REST*	0.911	0	0	0.928
IR/iGluR	IR	Dsil134293	386	*LEG#1, LEG#234*	0	0.178	0.207	0
IR/iGluR	AMPA	Dsil137319	217	*LEG#1, LEG#234*	0	0.407	0.471	0
IR/iGluR	AMPA	Dsil155434	276	*PALP, REST*	0.308	0	0	0.626
IR/iGluR	AMPA	Dsil37747	3,239	*PALP, LEG#1, LEG#234, REST*	0.135	0.212	0.306	0.499
IR/iGluR	AMPA	Dsil56426	1,742	*PALP, LEG#1, LEG#234, REST*	0.037	0.085	0.132	0.649
IR/iGluR	AMPA	Dsil7947	3,218	*PALP, LEG#1, LEG#234, REST*	0.197	0.034	0.264	30.515
IR/iGluR	AMPA	Dsil90016	518	*LEG#1, LEG#234, REST*	0	0.11	0.248	0.093
IR/iGluR	Kainate	Dsil10178	2,913	*PALP, LEG#1, LEG#234, REST*	0.825	1.492	1.521	0.545
IR/iGluR	Kainate	Dsil10448	3,652	*PALP, LEG#1, LEG#234, REST*	3.385	3.782	4.685	2.435
IR/iGluR	Kainate	Dsil1341	9,297	*PALP, LEG#1, LEG#234, REST*	13.268	29.256	32.08	13.669
IR/iGluR	Kainate	Dsil14755	4,305	*PALP, LEG#1, LEG#234, REST*	0.197	1.815	1.884	6.551
IR/iGluR	Kainate	Dsil23601	3,386	*PALP, LEG#1, LEG#234, REST*	0.295	0.729	0.835	1.785
IR/iGluR	Kainate	Dsil2489	3,499	*PALP, LEG#1, LEG#234, REST*	16.838	40.552	40.467	16.394
IR/iGluR	Kainate	Dsil2806	3,638	*PALP, LEG#1, LEG#234, REST*	13.921	27.272	17.155	10.179
IR/iGluR	Kainate	Dsil29949	5,707	*PALP, LEG#1, LEG#234, REST*	0.049	0.051	0.091	2.319
IR/iGluR	Kainate	Dsil4043	2,676	*PALP, LEG#1, LEG#234, REST*	0.308	0.144	0.744	81.922
IR/iGluR	Kainate	Dsil42393	3,253	*PALP, LEG#1, LEG#234, REST*	0.025	0.068	0.124	1.02
IR/iGluR	Kainate	Dsil56507	1,393	*PALP, LEG#1, LEG#234, REST*	0.185	0.136	0.066	0.638
IR/iGluR	Kainate	Dsil73459	731	*PALP, REST*	0.111	0	0	0.928
IR/iGluR	Kainate	Dsil73623	1,001	*PALP, LEG#1, LEG#234, REST*	0.037	0.042	0.025	0.881
IR/iGluR	Kainate	Dsil81800	824	*PALP, LEG#1, LEG#234, REST*	0.049	0.204	0.198	0.046
IR/iGluR	NMDA	Dsil136567	528	*REST*	0	0	0	0.696
IR/iGluR	NMDA	Dsil23576	3,395	*PALP, LEG#1, LEG#234, REST*	0.037	0.11	0.066	4.626
IR/iGluR	NMDA	Dsil25426	3,193	*PALP, LEG#1, LEG#234, REST*	0.935	0.56	0.62	0.603
IR/iGluR	NMDA	Dsil47451	1,211	*LEG#1, REST*	0	0.034	0	1.043
IR/iGluR	NMDA	Dsil53904	514	*REST*	0	0	0	1.542

Table S5C. Expression levels of transcripts encoding chemosensory genes across conditions

Chemosensory Gene family	Transcript	Length	Location	FPKM			
				PALP	*LEG#1*	*LEG#234*	*REST*
GR	Dsil110148	390	*PALP, LEG#1*	0.295	0.178	0	0
GR	Dsil118714	527	*LEG#1, REST*	0	0.051	0	0.522
GR	Dsil125602	286	*PALP, LEG#1, REST*	0.283	0.348	0	0.278
GR	Dsil126113	411	*LEG#1, REST*	0	0.322	0	0.128
GR	Dsil137841	300	*PALP, LEG#1*	0.258	0.305	0	0
GR	Dsil164676	201	*REST*	0	0	0	0.928
GR	Dsil32817	1,568	*PALP, LEG#1, LEG#234, REST*	0.271	0.136	0.058	2.748
GR	Dsil43425	1,557	*PALP, LEG#1, LEG#234, REST*	0.185	0.042	0.041	0.788
GR	Dsil46281	654	*PALP, LEG#1, LEG#234, REST*	0.295	0.076	0.256	0.278
GR	Dsil61916	635	*REST*	0	0	0	0.661
GR	Dsil64584	480	*PALP, LEG#1, LEG#234*	0.098	0.187	0.281	0
GR	Dsil85717	607	*PALP, LEG#1, LEG#234*	0.148	0.39	0.05	0

Table S5D. Expression levels of transcripts encoding chemosensory genes across conditions

Chemosensory Gene family	Transcript	Length	Location	FPKM			
				PALP	*LEG#1*	*LEG#234*	*REST*
OBP-like	Dsil1642	1,228	*PALP, LEG#1, LEG#234, REST*	95.673	43.308	36.013	14.736
OBP-like	Dsil553	1,730	*PALP, LEG#1, LEG#234, REST*	33.282	78.848	89.182	629.627
OBP-like	Dsil709	1,160	*PALP, LEG#1, LEG#234, REST*	222.782	113.938	111.766	96.913

Table S5E. Expression levels of transcripts encoding chemosensory genes across conditions

Chemosensory Gene family	Transcript	Length	Location	FPKM			
				PALP	*LEG#1*	*LEG#234*	*REST*
NPC2	Dsil113431	330	*LEG#1*	0	0.127	0	0
NPC2	Dsil1152	938	*PALP, LEG#1, LEG#234, REST*	3.939	1.408	5.057	492.692
NPC2	Dsil13162	727	*PALP, LEG#1, LEG#234, REST*	3.976	4.477	1.661	4.626
NPC2	Dsil16636	650	*PALP, LEG#1, LEG#234, REST*	8.37	0.076	0.091	0.707
NPC2	Dsil2309	762	*PALP, LEG#1, LEG#234, REST*	6.45	0.856	3.793	134.917
NPC2	Dsil28573	705	*PALP, LEG#1, LEG#234, REST*	1.415	0.594	0.587	1.264
NPC2	Dsil56450	1,171	*PALP, LEG#1, LEG#234, REST*	4.923	3.307	3.363	0.464
NPC2	Dsil6690	688	*PALP, LEG#1, LEG#234, REST*	2.56	0.399	1.421	50.132
NPC2	Dsil700	1,515	*PALP, LEG#1, LEG#234, REST*	53.529	27.611	23.907	343.838
NPC2	Dsil793	7,327	*PALP, LEG#1, LEG#234, REST*	50.575	40.153	27.948	23.373
NPC2	Dsil93094	273	*PALP, REST*	1.292	0	0	0.325

Table S5F. Expression levels of transcripts encoding chemosensory genes across conditions

Chemosensory Gene family	Transcript	Length	Location	FPKM			
				PALP	*LEG#1*	*LEG#234*	*REST*
CD36	Dsil1051	2,035	*PALP, LEG#1, LEG#234, REST*	98.258	59.801	49.268	70.189
CD36	Dsil115517	420	*LEG#234, REST*	0	0	0.083	0.128
CD36	Dsil12516	2,326	*PALP, LEG#1, LEG#234, REST*	1.563	0.178	0.719	8.556
CD36-SNMPs	Dsil16834	1,400	*PALP, LEG#1, LEG#234, REST*	2.191	2.145	1.843	1.206
CD36-SNMPs	Dsil1909	12,941	*PALP, LEG#1, LEG#234, REST*	62.711	32.275	29.063	51.871
CD36	Dsil2138	1,839	*PALP, LEG#1, LEG#234, REST*	45.96	45.648	29.518	28.138
CD36-SNMPs	Dsil31189	1,732	*PALP, LEG#1, LEG#234, REST*	1.785	1.179	2.553	0.441
CD36	Dsil3182	2,809	*PALP, LEG#1, LEG#234, REST*	0.714	0.212	0.719	64.485
CD36	Dsil42334	1,038	*PALP, LEG#1, LEG#234, REST*	0.862	0.212	0.397	0.313
CD36	Dsil43685	252	*LEG#1, LEG#234, REST*	0	0.492	0.562	6.052
CD36	Dsil51237	1,854	*PALP, LEG#1, LEG#234, REST*	0.148	0.051	0.058	0.73
CD36	Dsil6227	2,130	*PALP, LEG#1, LEG#234, REST*	3.963	4.104	4.363	20.718
CD36-SNMPs	Dsil935	4,957	*PALP, LEG#1, LEG#234, REST*	44.643	50.575	41.88	91.406

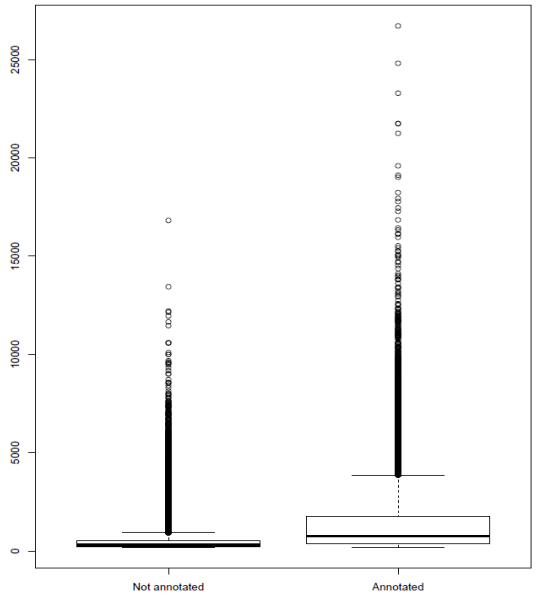
Table S5G. Expression levels of transcripts encoding chemosensory genes across conditions

Chemosensory Gene family	Transcript	Length	Location	FPKM			
				PALP	*LEG#1*	*LEG#234*	*REST*
CCP	Dsil1144	646	*PALP, LEG#1, LEG#234, REST*	36.088	29.146	11.924	0.58
CCP	Dsil11975	480	*PALP, LEG#1, LEG#234, REST*	0.209	12.567	2.768	0.499
CCP	Dsil1773	1,127	*PALP, LEG#1, LEG#234, REST*	109.582	118.848	37.005	0.754
CCP	Dsil1894	278	*PALP, LEG#1, LEG#234, REST*	69.013	51.983	23.378	0.301
CCP	Dsil27094	868	*PALP, LEG#1, LEG#234*	2.806	3.748	2.694	0
CCP	Dsil2713	363	*PALP, LEG#1, LEG#234, REST*	50.144	69.799	26.237	0.162
CCP	Dsil329	2,280	*PALP, LEG#1, LEG#234, REST*	202.239	156.381	26.312	2.759
CCP	Dsil44527	669	*PALP, LEG#1, LEG#234*	0.382	0.305	0.215	0
CCP	Dsil502	785	*PALP, LEG#1, LEG#234, REST*	194.472	265.79	108.659	1.646
CCP	Dsil516	1,904	*PALP, LEG#1, LEG#234, REST*	63.548	73.844	27.725	0.835
CCP	Dsil562	626	*PALP, LEG#1, LEG#234*	17.995	21.048	7.685	0
CCP	Dsil809	330	*PALP, LEG#1, LEG#234, REST*	159.96	131.958	47.045	17.182

Table S6. Relationship between functional annotation success and transcript length

Length Range	Contigs*	Transcripts	Annotated transcripts	% of Annotated transcripts
<300	67,147	66,721	7,263	10.89%
301-600	64,422	54,976	9,886	17.98%
601-1000	35,022	21,568	6,577	30.49%
1001-2000	35,789	16,073	8,245	51.30%
>2000	33,903	11,508	9,075	78.86%
Total	236,283	170,846	41,046	24.03%

*Non-clustered transcripts



The statistical significance of the association between transcript length and functional annotation was tested with an ANOVA one-way test (P-value = $2.2x10^{-16}$). The test was conducted by tabulating all transcripts in two groups, with annotated or not-annotated information (represented as a boxplot)

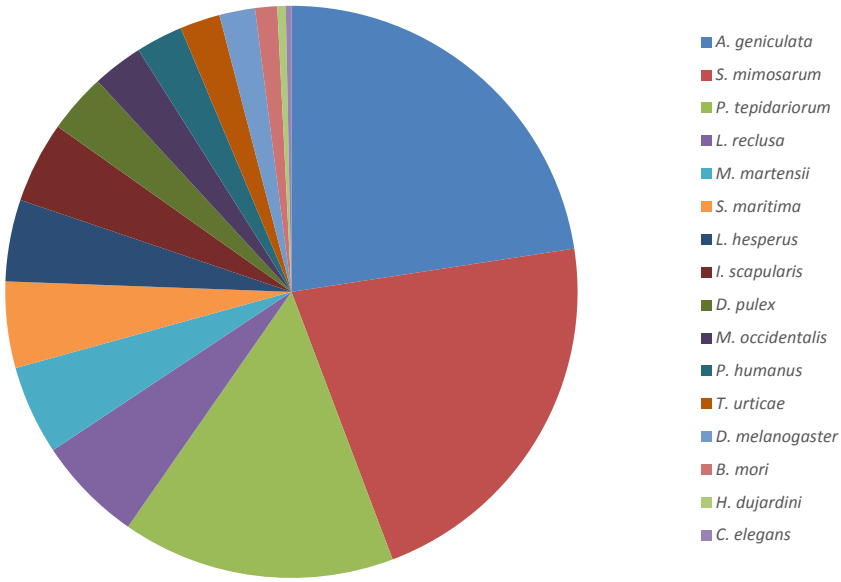


Figure S1. Distribution of blastx hits across species. Distribution of the top 5 hits from the blastx search of the 170,846 *D. silvatica* transcripts against the arthropodDB database with an *E*-value lower than 10^{-3}.

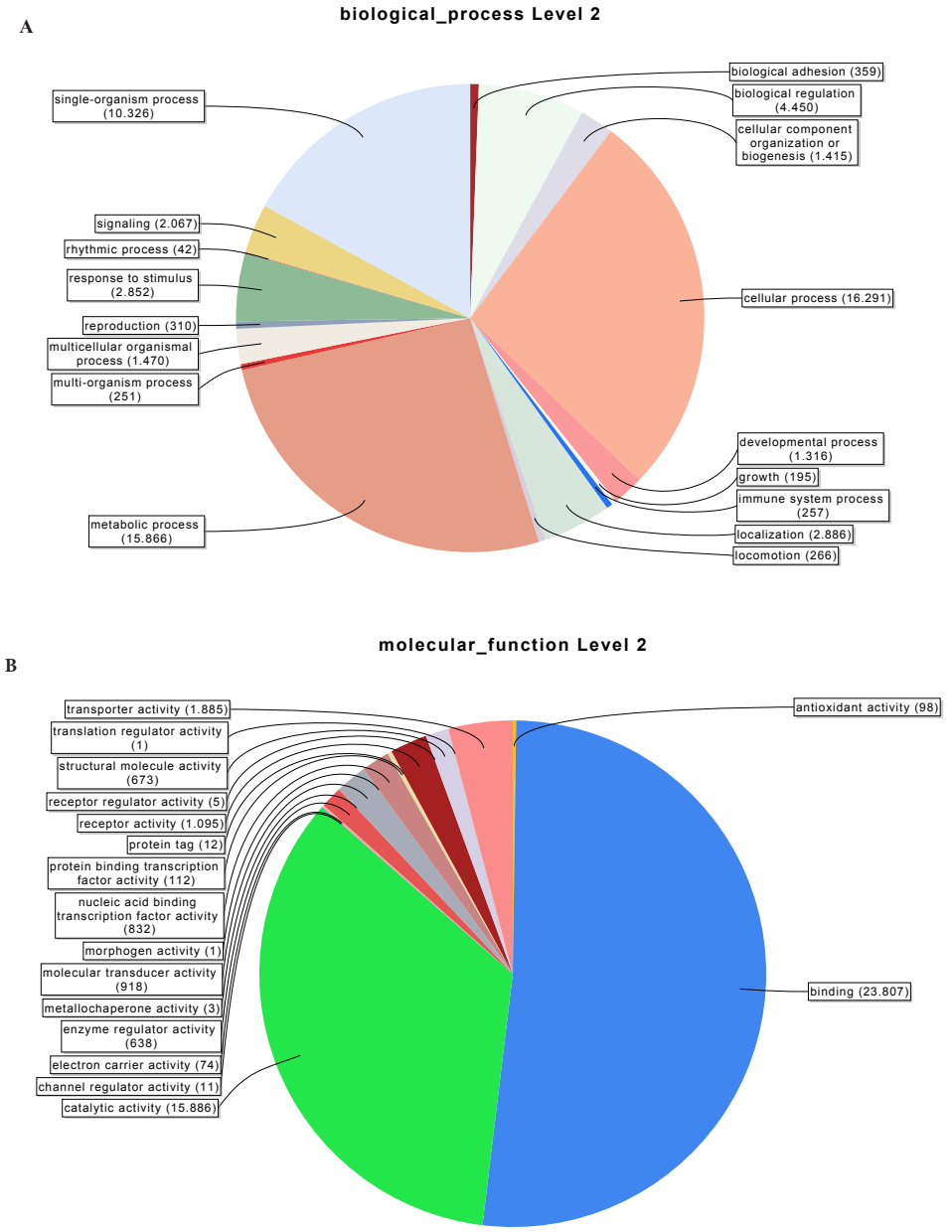


Figure S2. Distribution of the Gene Ontology (GO) terms associated with *D. silvatica* transcripts (29,879 transcripts with GO annotation from a total of 170,846 transcripts). (*A*) Biological process terms. (*B*) Molecular function terms.

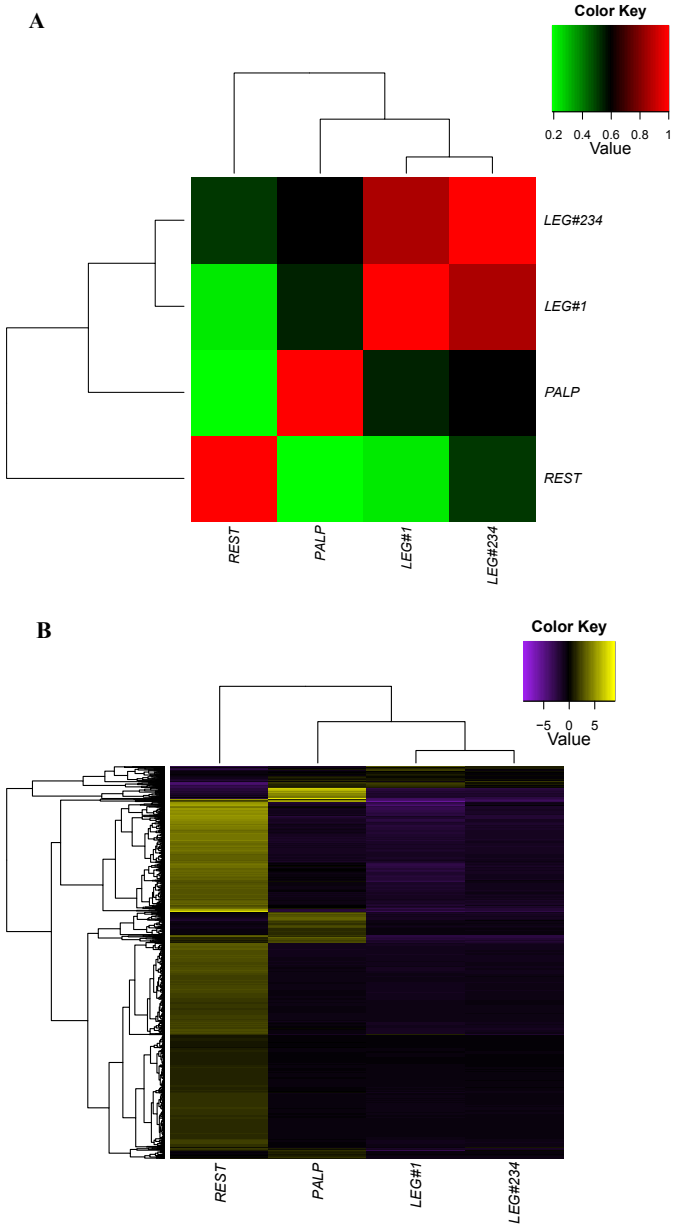


Figure S3. Heat maps reflecting the gene expression profiles of the four experimental conditions based on the 2,964 differentially expressed genes (P-value $\leq 10^{-3}$). (*A*) Comparison between conditions. (*B*) Expression profile of each gene across conditions.

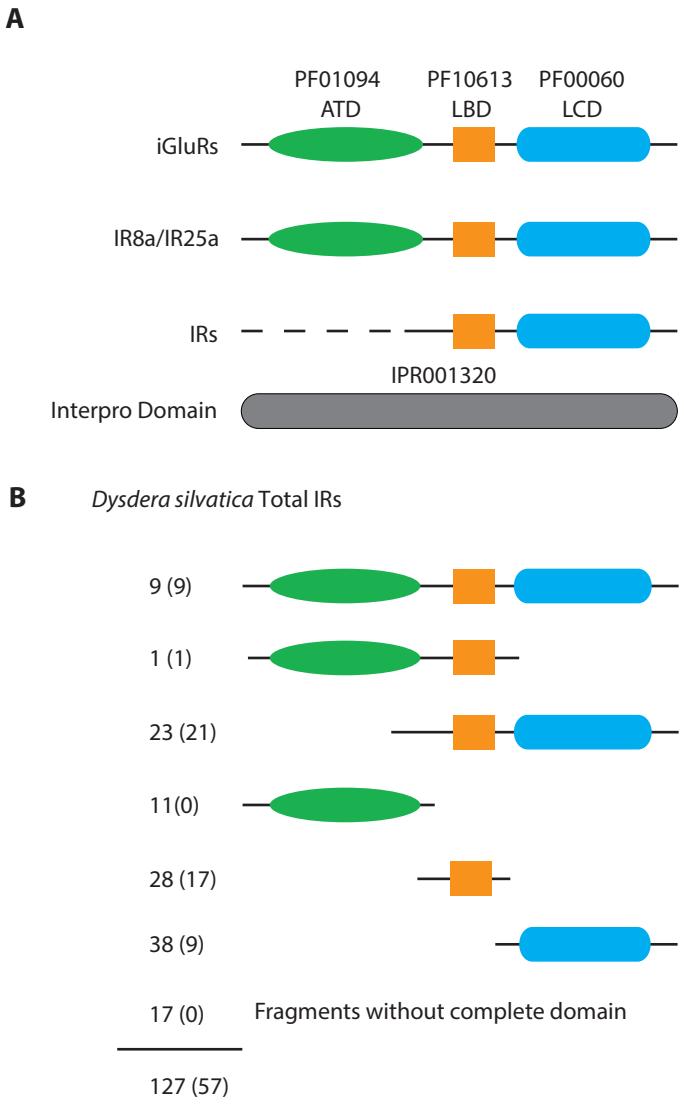


Figure S4. Protein domain organization of the IR/iGluR family proteins. (*A*) Domain organization of the different IR/iGluR subfamilies. (*B*) Distribution of the known IR/iGluR domains in the 127 transcripts characterized in *D. silvatica*. The numbers in parentheses show those transcripts exhibiting the InterPro domain.

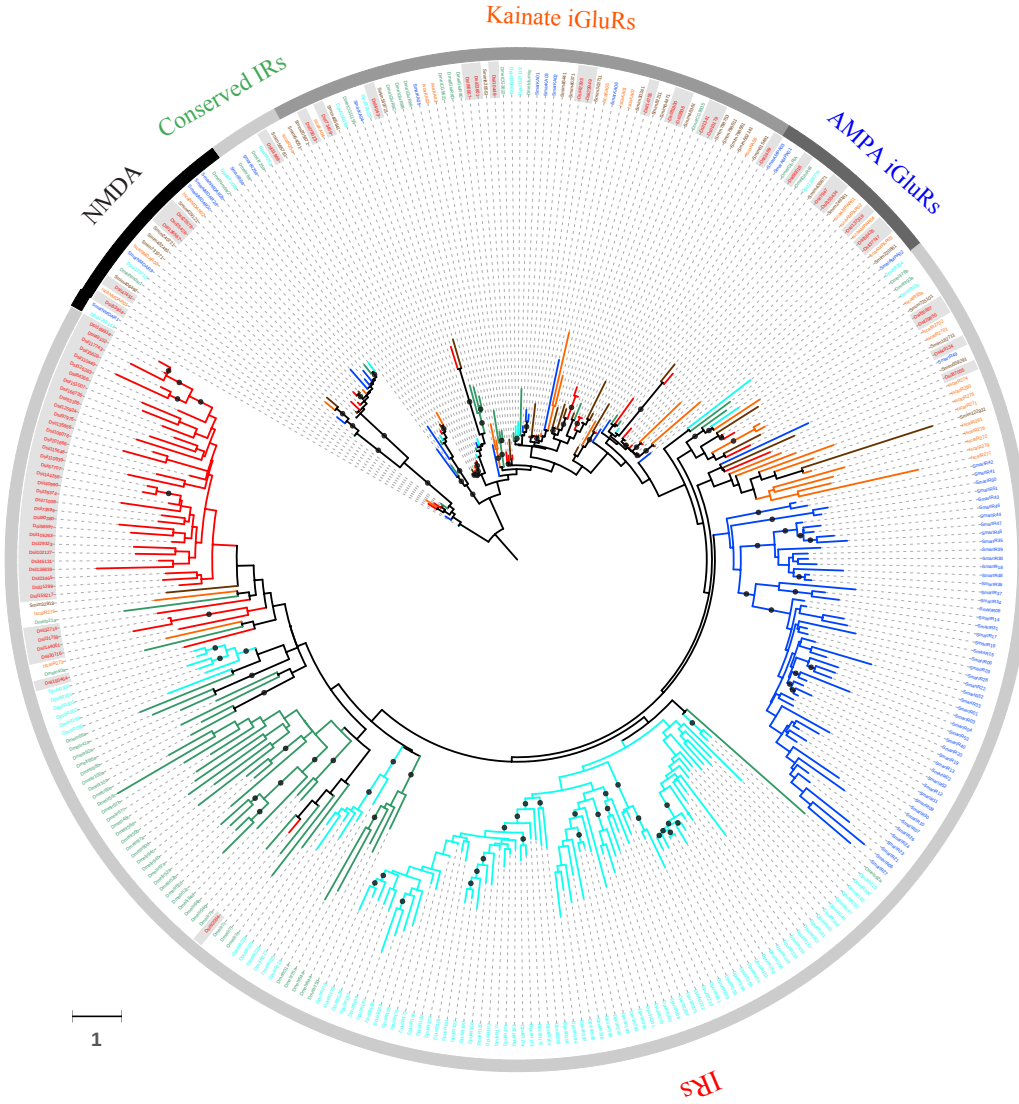


Figure S5. Maximum likelihood phylogenetic tree of the IR/iGluR proteins across arthropods. This un-collapsed phylogenetic tree includes the same protein sequences as in Fig. 3.

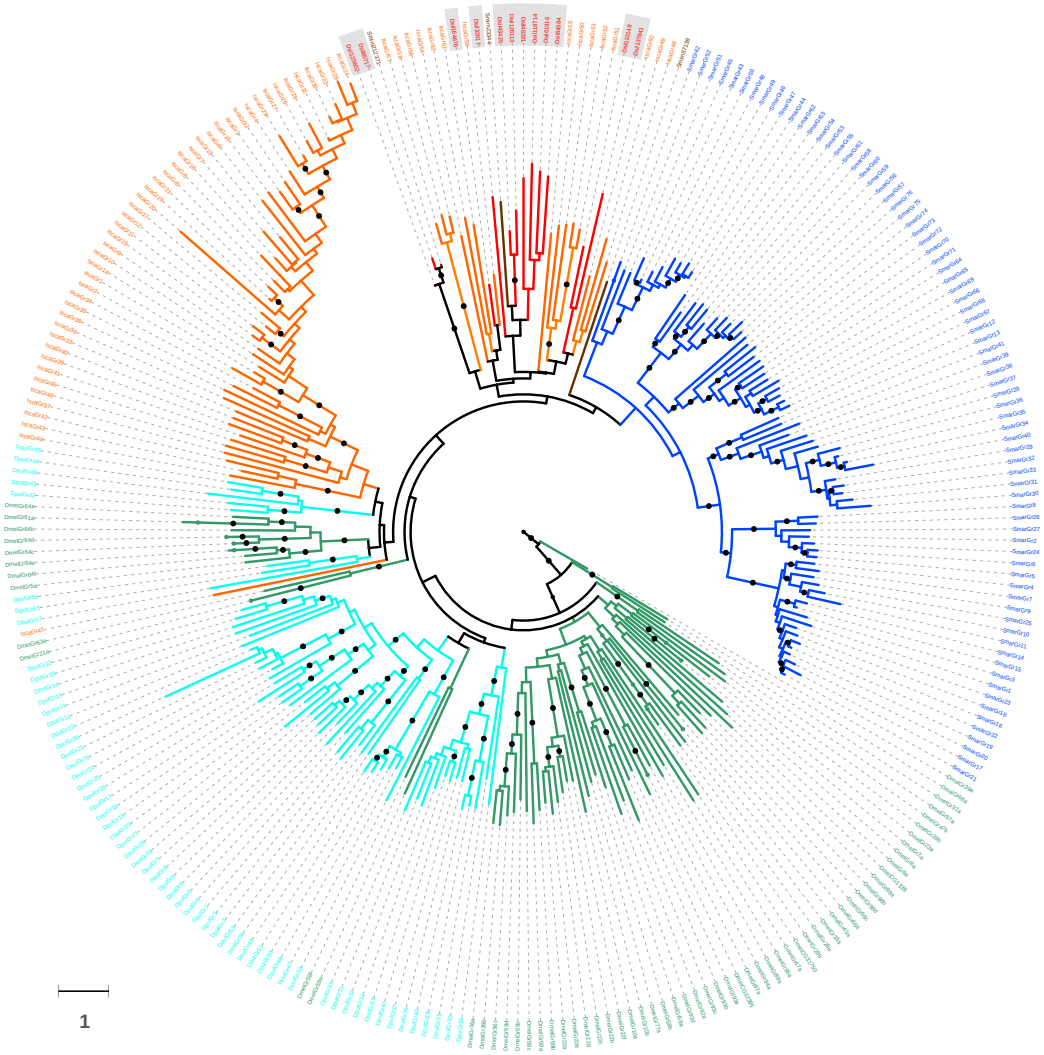


Figure S6. Maximum likelihood phylogenetic tree of the GR proteins across arthropods. This un-collapsed phylogenetic tree includes the same protein sequences as in Fig. 4.

Figure S7. MSAs of the nine OBP-*like* proteins and all characterized members of the *Obp* family in *D. melanogaster* and *A. gambiae.* (*A*) MAFFT-based alignment. (*B*) PROMAL3D-based alignment. (*C*) PSI-Coffee-based alignment. Available at *Genome Biology and Evolution* online https://doi.org/10.1093/gbe/evw296

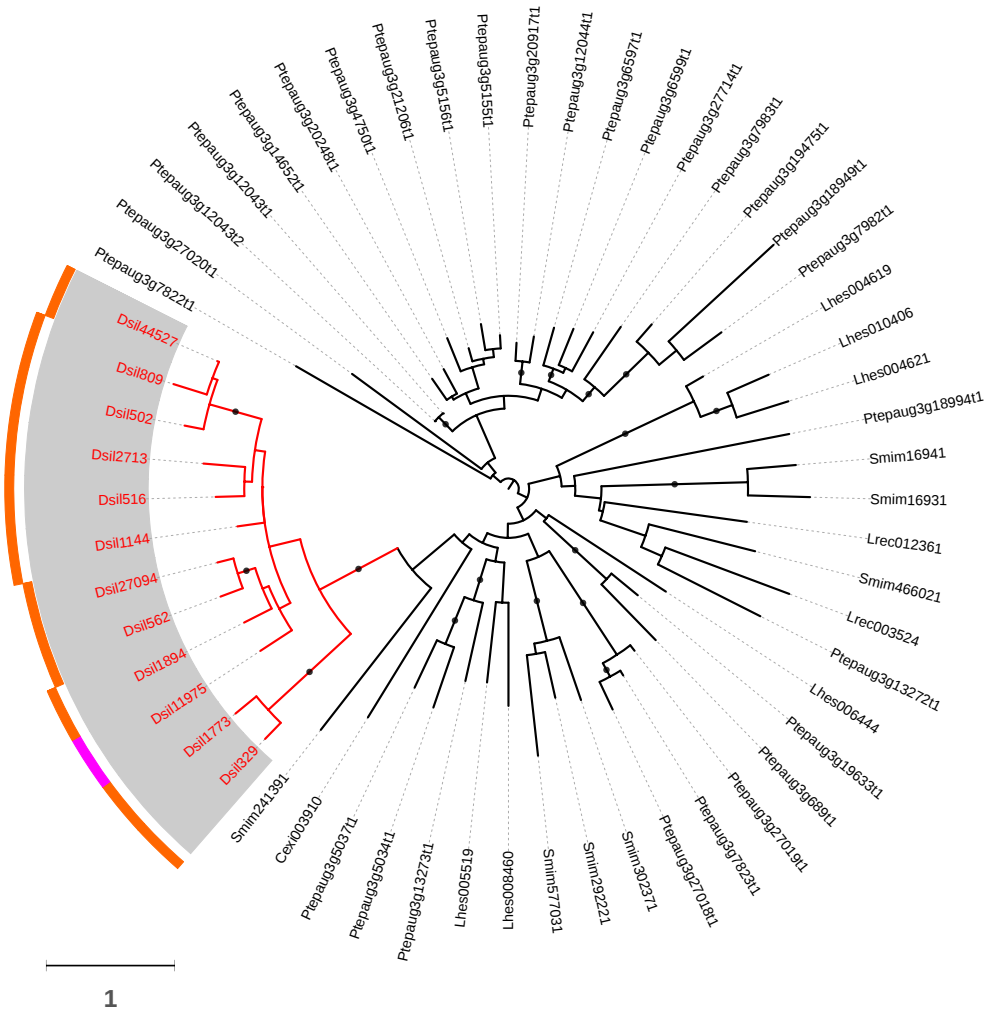


Figure S8. Maximum likelihood phylogenetic tree of the CCPs found in chelicerates. Lhes, *L. hesperus*; Cexi, *C. exilicauda*; Ptep, *P. tepidariorum*; Lrec, *L. reclusa*; Smim, *S. mimosarum*; Dsil, *D. silvatica*. The translations of the *D. silvatica* Ccp transcripts are colored in red. Node support features and surrounding circles are colored as in Fig. 3.

3

Comparative genomics reveals thousands of novel chemosensory genes and massive changes in chemoreceptor repertories across chelicerates

La quimiopercepción es una función biológica esencial para la supervivencia, reproducción, e incluso la comunicación social de los animales. A pesar de que los mecanismos moleculares involucrados en la quimiopercepción en insectos son relativamente bien conocidos, existen pocos estudios en otros linajes de artrópodos. La disponibilidad actual de genomas de un gran número de quelicerados ofrece la oportunidad de caracterizar las familias multigénicas involucradas en esta función en un linaje que se originó y colonizó el medio de terrestre de forma independiente a los insectos. A su vez, esto ofrece nuevas oportunidades y retos para el estudio de este linaje animal en muchas áreas de investigación. En este trabajo, hemos realizado un estudio genómico comparativo teniendo en consideración la alta fragmentación en los genomas disponibles, y, por primera vez, incluyendo datos genómicos de especies que cubren la mayoría de diversidad de quelicerados. Nuestras búsquedas exhaustivas identificaron miles de genes quimiosensoriales previamente no caracterizados, la mayoría de ellos codificando quimiorreceptores ionótropicos y gustativos (IRs y GRs). El análisis filogenético y de ganancia y pérdida de genes indican que los eventos de duplicaciones genómicos globales propuestos en este subfilo no explicarían las diferencias en el repertorio de quimiorreceptores observado entre especies. El proceso de nacimiento y muerte de genes, influido por duplicaciones génicas episódicas originando expansiones específicas de linaje, también habría contribuido de forma importante a la diversidad existente en estas familias quimiosensoriales. Este estudio también profundiza en el origen y evolución de otras familias de genes quimiosensoriales diferentes a los receptores, como las OBPs (*odorant-binding proteins*) y otras proteínas relacionadas con la quimiopercepción.

GBE

Comparative Genomics Reveals Thousands of Novel Chemosensory Genes and Massive Changes in Chemoreceptor Repertories across Chelicerates

Joel Vizueta, Julio Rozas*, and Alejandro Sánchez-Gracia*

Departament de Genètica, Microbiologia i Estadística and Institut de Recerca de la Biodiversitat (IRBio), Facultat de Biologia, Universitat de Barcelona, Barcelona, Spain

*Corresponding authors: E-mails: jrozas@ub.edu;elsanchez@ub.edu.

Accepted: April 17, 2018

Data deposition: All data generated or analyzed during this study are included in this published article (and its supplementary file, Supplementary Material online).

Abstract

Chemoreception is a widespread biological function that is essential for the survival, reproduction, and social communication of animals. Though the molecular mechanisms underlying chemoreception are relatively well known in insects, they are poorly studied in the other major arthropod lineages. Current availability of a number of chelicerate genomes constitutes a great opportunity to better characterize gene families involved in this important function in a lineage that emerged and colonized land independently of insects. At the same time, that offers new opportunities and challenges for the study of this interesting animal branch in many translational research areas. Here, we have performed a comprehensive comparative genomics study that explicitly considers the high fragmentation of available draft genomes and that for the first time included complete genome data that cover most of the chelicerate diversity. Our exhaustive searches exposed thousands of previously uncharacterized chemosensory sequences, most of them encoding members of the gustatory and ionotropic receptor families. The phylogenetic and gene turnover analyses of these sequences indicated that the whole-genome duplication events proposed for this subphylum would not explain the differences in the number of chemoreceptors observed across species. A constant and prolonged gene birth and death process, altered by episodic bursts of gene duplication yielding lineage-specific expansions, has contributed significantly to the extant chemosensory diversity in this group of animals. This study also provides valuable insights into the origin and functional diversification of other relevant chemosensory gene families different from receptors, such as odorant-binding proteins and other related molecules.

Key words: chemosensory gene family, gustatory receptors, ionotropic receptors, acari, spiders, scorpions.

Introduction

The i5k initiative (Robinson et al. 2011) has greatly boosted the complete genome sequencing and functional annotation of a number of arthropod species. The currently available genome data were obtained from species chosen for their significance as model organisms in diverse areas, such as agriculture, medicine, food safety or biodiversity, or for their strategic phylogenetic position in evolutionary studies on the diversification of the major arthropod lineages (Adams et al. 2000; Colbourne et al. 2011; Cao et al. 2013; Chipman et al. 2014; Sanggaard et al. 2014; Gulia-Nuss et al. 2016). As expected, the first sequencing initiatives focused on insects, although the number of sequenced noninsect genomes has increased considerably over time, especially in chelicerates. The recent genome sequence data from chelicerate species (Cao et al. 2013; Sanggaard et al. 2014; Gulia-Nuss et al. 2016) are disrupting the strongly biased taxonomic distribution of arthropod genomes hitherto available. More importantly, these new data have greatly facilitated studies on the origin and evolutionary divergence of this highly diverse animal subphylum (Kenny et al. 2016; Schwager et al. 2017), which has important impacts on translational research such as silk production in spiders, biomedical applications of spider and scorpion venom toxins, or plague control in acari

(Mille et al. 2015; Hoy et al. 2016; Babb et al. 2017; Gendreau et al. 2017; Pennisi 2017).

Chemoreception is a paradigmatic example of a relatively well-known biological system in insects, but it is not as well characterized in other arthropods despite numerous practical applications as pest control strategies, biosensors or electronic nose sensors (Berna et al. 2009; Wei et al. 2017). In chelicerates, as in other animals, the chemosensory system (CS) is critical for the survival, reproduction, and social communication of individuals. The detection and integration of environmental chemical signals, including smell and taste, allow organisms to detect food, hosts, and predators and frequently play a crucial role in social communication (Joseph and Carlson 2015). In *Drosophila*, peripheral events occur in specialized hair-like cuticular structures (sesilla) that are distributed throughout the body surface, with a prominent concentration in antennae and maxillary palps (olfactory sensilla) or on the distal tarsal segments of the legs (gustatory sensilla) (Pelosi 1996; Shanbhag et al. 2001). In this species, chemoreceptor proteins, which are located in the membranes of sensory neurons innervating the sensillum lymph, convert the external chemical signal into an electrical one, which is, in turn, processed in higher brain regions (de Bruyne and Baker 2008; Sánchez-Gracia et al. 2009; Sato and Touhara 2009). The sensillum lymph contains a set of highly abundant small globular proteins (hereafter termed "binding proteins") that are thought to bind to, solubilize and transport chemical cues to the space surrounding chemoreceptors (Vogt and Riddiford 1981; Pelosi et al. 2006). The genome of the fruit fly encodes two different kinds of membrane chemoreceptors that are phylogenetically unrelated. The first group comprises the superfamily of insect olfactory (*Or*) and gustatory (*Gr*) receptors, which encode seven-transmembrane receptors with an atypical membrane topology and heteromeric function, and share a common origin (Missbach et al. 2015). Interestingly, and despite performing analogous functions, these receptors are structurally and genetically unrelated to their vertebrate counterparts, where G protein-coupled receptors are involved in chemoreception (Kaupp 2010). The second group of chemoreceptors encodes the ionotropic receptor (*Ir*) gene family, a highly divergent lineage that is related to the ionotropic glutamate receptors superfamily (*iGluR*) associated with both olfaction and taste functions (Robertson and Wanner 2006; Benton et al. 2009; He et al. 2013; Missbach et al. 2014). The extracellular binding proteins of *Drosophila* include the odorant binding protein (*Obp*), chemosensory protein (*Csp*), chemosensory proteins A and B (*CheA* and *CheB*) and Niemann–Pick Type C2 (*Npc2*) families (Li et al. 2008; Dani et al. 2011; Iovinella et al. 2011). Moreover, sensory neuron membrane proteins (SNMPs), which are related to the CD36 receptor family and expressed in specific *Drosophila* pheromone-responding sensory neurons, also play a key role in sensory perception by facilitating the contact between ligand and receptor (Gomez-Diaz et al. 2016). It is worth noting that there is a lack of evidence that all CS family members actually possess a true chemosensory function, and they are usually classified as chemosensory-related genes based on their sequence similarity with previously examined members (Kitabayashi et al. 1998; Wanner et al. 2005; Ishida et al. 2013; Joseph and Carlson 2015).

There are few comprehensive studies of the characterization and classification of CS gene families in noninsect genomes, with only six noninsect arthropod species investigated to date: The crustacean *Daphnia pulex*, the myriapods *Strigamia maritima* and *Trigoniulus corallinus*, and the chelicerates *Ixodes scapularis*, *Metaseiulus occidentalis* and *Tetranychus urticae* (Colbourne et al. 2011; Chipman et al. 2014; Kenny et al. 2015; Gulia-Nuss et al. 2016; Hoy et al. 2016; Ngoc et al. 2016). Moreover, we and others have also reported transcriptome data for various chelicerate species (Frías-López et al. 2015; Qu et al. 2016; Eliash et al. 2017; Vizueta et al. 2017). These works confirm that chelicerates contain members of all insect CS gene families, with the single exception of the *Or* family (Benton et al. 2007; et al. 2011; Missbach et al. 2014), which likely emerged from a *Gr* ancestor during the diversification of flying insects (Missbach et al. 2015). The recent identification of two novel candidate CS families in chelicerates, the *Obp*-like and the candidate carrier protein (*Cpp*) families, is also remarkable (Vizueta et al. 2017). The *Obp*-like family, which encodes proteins with some sequence and structural similarity to canonical insect OBPs, has also been identified in centipedes (Vizueta et al. 2017), and this finding makes unclear the evolution of these gene families in arthropods. The *Ccp* family, which was first discovered in the transcriptome of *D. silvatica*, contains members that are differentially expressed in the putative chemosensory appendages of this spider. Although OBP-like and CCPs share some common structural features with other CS proteins, their potential functional roles as chemosensory proteins and the extent to which these proteins are present in arthropods remain to be elucidated (Renthal et al. 2017; Vizueta et al. 2017).

The ancestor of all extant chelicerates can be traced back to the Cambrian period (~530 Ma); therefore, this group colonized land independently of the other arthropod lineages (*Hexapoda*, *Crustacea*, and *Myriapoda*; Rota-Stabelli et al. 2013). As there are no OR-encoding genes, other proteins likely perform OR's function. Current experimental data from noninsect arthropods, such as the specific gene expression and electrophysiological recording data for some IR members in the olfactory structures of lobsters and hermit crabs (Corey et al. 2013; Groh-Lunow et al. 2015) and RNA-seq of the palps and first pair of legs of spiders (Vizueta et al. 2017) and centipede antennas (C. Frias-López, F.C. Almeida, S. Guirao-Rico, R. Jenner, A. Sánchez-Gracia and J. Rozas, unpublished results), indicate that this receptor family contains the best candidates for actual olfactory receptors. The specific organs and molecules responsible for gustatory function are less well understood; nevertheless, as some *Gr* and *Ir* family members are

differentially expressed across some body parts in these species, contact chemoreceptors appear to be the best candidates. Given this difference in functional roles of the various CS families, it is highly relevant to gain further comprehensive insights into their evolution in arthropods other than insects/hexapods.

Here, we carried out an enhanced comparative genomic analysis of the CS families across 11 chelicerate genomes. We applied powerful sequence similarity-based searches using state-of-the-art methodologies and expressly considered the fragmented nature of the surveyed genomes. We conducted a comprehensive phylogenetic analysis of chemosensory genes from different gene families and characterized the turnover rates of chemoreceptor families across chelicerates after accurate estimation of the number of gene duplications and gene losses in each lineage. We also contribute new knowledge about some interesting questions that are not yet fully resolved, such as the evolutionary relationship between OBP and OBP-like proteins or the extent in which CCP and CSP are present in chelicerates.

Materials and Methods

Genomic Data

We retrieved all genomic sequences, annotations, and predicted peptides of 14 arthropods, including 11 chelicerates, from public databases (fig. 1). Specifically, we used the genome information of the fruit fly *Drosophila melanogaster* (r6.05, FlyBase) (Adams et al. 2000), the crustacean *Daphnia pulex* (r1.26, Ensembl Genomes) (Colbourne et al. 2011), and the centipede *Strigamia maritima* (r1.26, Ensembl Genomes) (Chipman et al. 2014). The chelicerate genomes included the horseshoe crab *Limulus polyphemus* (v2.1.2, NCBI Genomes) (Nossa et al. 2014); the acari *Tetranychus urticae* (r1.26, Ensembl Genomes) (Grbić et al. 2011), *Metaseiulus occidentalis* (v1.0, NCBI Genomes) (Hoy et al. 2016), and *Ixodes scapularis* (r1.26, Ensembl Genomes) (Gulia-Nuss et al. 2016); the scorpions *Centruroides exilicauda* (bark scorpion, genome assembly version v1.0, annotation version v0.5.3; Human Genome Sequencing Center [HGSC]) and *Mesobuthus martensii* (v1.0, Scientific Data Sharing Platform Bioinformation [SDSPB]; Cao et al. 2013); and the spiders *Acanthoscurria geniculata* (tarantula, v1, NCBI Assembly, BGI; Sanggaard et al. 2014), *Stegodyphus mimosarum* (African social velvet spider, v1, NCBI Assembly, BGI; Sanggaard et al. 2014), *Latrodectus hesperus* (western black widow, v1.0, HGSC), *Parasteatoda tepidariorum* (common house spider, v1.0 Augustus 3, SpiderWeb and HGSC; Schwager et al. 2017), and *Loxosceles reclusa* (brown recluse, v1.0, HGSC).

Query Data Sets and Protein Search Protocol

Our comprehensive CS search protocol included the creation of three data sets, which were iteratively used as queries in successive hierarchical rounds of sequence similarity- and profile-based searches (fig. 2).

Data Set 1

The starting data set contained the CS proteins from publicly available, well-annotated genomes. This data set included the protein sequences of the *Gr*, *Ir*/*iGluR*, *Or*, *Csp*, *Obp*, *Npc2*, and *Snmp-Cd36* families from 1) the hexapods *D. melanogaster* (Benton et al. 2009; Vogt et al. 2009; Vieira and Rozas 2011; Pelosi et al. 2014), *T. castaneum* (Sánchez-Gracia et al. 2009; Croset et al. 2010; Dippel et al. 2014), *A. pisum* (Zhou et al. 2010), and *A. mellifera* (Robertson and Wanner 2006; Forêt et al. 2007; Nichols and Vogt 2008); 2) the crustacean *D. pulex* (Peñalva-Arana et al. 2009); 3) the myriapod *S. maritima* (Chipman et al. 2014); and 4) the ticks *I. scapularis* (Gulia-Nuss et al. 2016), *M. occidentalis* (Hoy et al. 2016), and *T. urticae* (Ngoc et al. 2016).

Data Set 2

This data set included the sequences of data set 1 (DS1) plus the new identified CS protein sequences with specific CS protein domains (see Table S1 in Vizueta et al. [2017] for details). We applied InterProScan (5.4.47; Jones et al. 2014) against genome-wide predicted peptides without a functional chemosensory annotation (i.e., in chelicerate genomes that were not used in the step to build DS1). Furthermore, we also included in data set 2 (DS2) the members of the *Cpp* family identified in Vizueta et al. (2017), as well as those found in current chelicerate genomes, after conducting several rounds of BlastP searches (version 2.2.30; Altschul 1997).

Data Set 3

This data set resulted from incorporating some additional highly curated sequences (a second search round against all surveyed genomes) into DS2. For that, we built for each CS family a multiple sequence alignment (MSA) of all DS2 proteins and the corresponding Pfam profile as a guide (using the HMMER software; Eddy 2011). We used these MSAs to build new (more specific) HMM profiles, with one per gene family (generically named CS-F-HMM). For the second search round of predicted peptides from all genomes, we used as queries both the CS-F-HMM profiles (in HMMER searches; i-*E*-value $< 10^{-5}$) and the sequences of DS2 (in BlastP searches; *E*-value $< 10^{-5}$). Moreover, we only retained the BlastP-positive hits for which the alignment between the query and the subject either covered at least two-thirds of the query length or included at least 80% of the subject peptide. Finally, we trimmed all the fragments not aligned between queries and the subject sequences and added the alignment region to DS2 to build data set 3 (DS3).

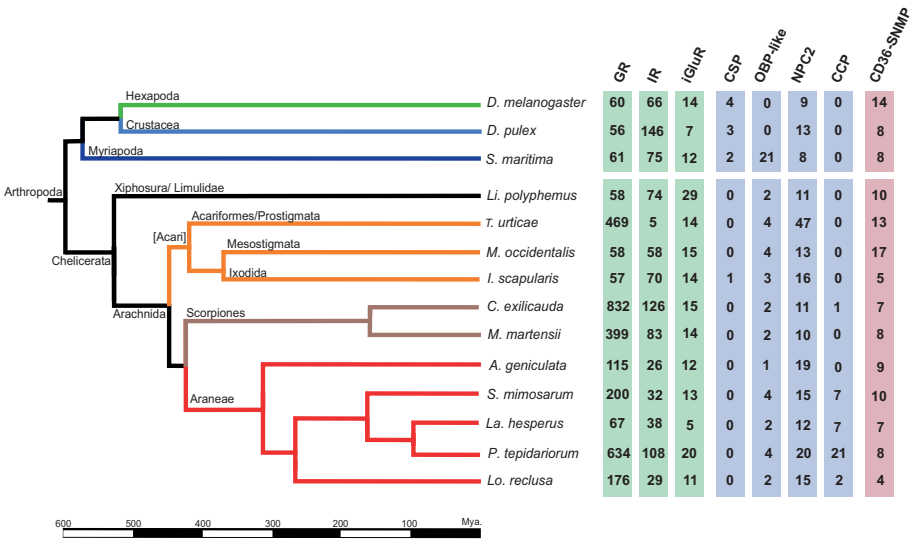


Fig. 1.—Phylogenetic relationships among the 14 surveyed species. Divergence times are given in millions of years. Some branches representative of major lineages are shaded in different colors. Green, insects; light blue, crustaceans; dark blue, myriapods; black, horseshoe crabs; orange, acari; brown, scorpions; red, spiders. Numbers in the right part of the figure indicate the number of CS encoding sequences separated per each family (S_{MIN} values).

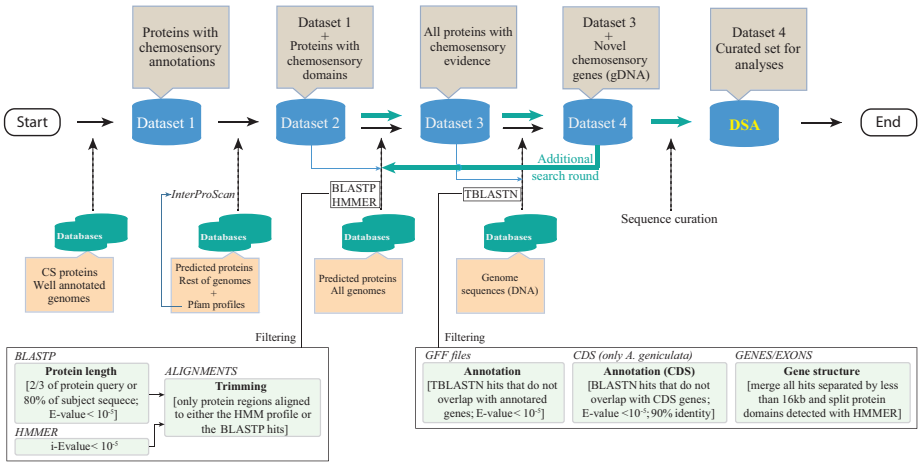


Fig. 2.—Workflow showing the steps used for the identification and annotation of the chemosensory gene families.

Data Set 4 and Data Set for the Analyses

Data set 4 (DS4) is the most curated and inclusive data set used for searches. The new information in DS4 was obtained after conducting exhaustive searches for CS-encoding regions directly on the DNA genome sequences using DS3 peptides as queries in a TBlastN search (*E*-value $< 10^{-5}$). Positive blast hits on regions that were not annotated in the GFF files were considered putative novel CS family members. For the genome of *A. geniculata*, where there is no GFF information,

we checked for the presence of any protein-coding region in the available transcriptomic data.

The TBlastN search allowed essentially the identification of exonic regions. To expand these regions to cover complete genes (as much as possible), we concatenated all sequences with hits located in the same scaffold and separated by <16 kb. We chose a 16-kb cut-off value because it corresponds to the 95th percentile of the intron length distribution in the studied genomes (i.e., fragments separated by higher distances are unlikely to be exons of the same gene). Next, we translated the nucleotide sequences according to the TBlastN reading frame. To avoid generating chimeric proteins from physically close but different genes, we used the specific CS-F-HMM profile to determine whether the number of different domains of each new protein after concatenation was compatible with a single gene (HMMER search; i-*E*-value $< 10^{-5}$). In addition to the "16-kb cut-off approach," and to try to extend a putative incomplete gene because the putative exons might be located in different scaffolds, we also applied the ESPRIT algorithm (Dessimoz et al. 2011) to join these partial fragments using DS1 as a guide. Finally, all the newly discovered CS-encoding sequences were added to DS3 to generate DS4. These protein data in DS4 were then used as a query to conduct an additional search round (in the same way as in the DS3 and DS4 steps). Finally, we conducted a semiautomatic step to curate the newly identified sequences from putative errors introduced in the search process (deletion of putative artefactual stop codons generated by TBlastN searches, splitting different genes erroneously fused in the same sequence, removing very small fragments). With the curated data, we established the final chelicerate CS protein data set, named DSA (data set for the analyses), which was used in further comparative genomic and evolutionary analyses (supplementary table S1*A*, Supplementary Material online). All new CS-proteins (including incomplete fragments) identified in this study are provided in the supplementary material, Supplementary Material online.

Functional and Structural Classification of CS Sequences

We classified the novel sequences in different categories based on structural and functional criteria. First, we examined the presence of premature stop codons; these features could represent real nonfunctional copies (pseudogenes), errors in sequencing or genome assembly steps or inaccuracies in our automatic annotation step based on TBlastN hits. All sequences encoding complete proteins (CPs) that were free of stop codons were included in the first category (CP set). Operationally, we considered a CP when its length was >80% of the corresponding average protein domain length. In addition, and only for the GR family, we also required that the CP members contained a minimum of 5 of the 7 transmembrane domains (defined by the software TMHMM version 2.0c; Krogh et al. 2001; Phobius version 1.01; Käll et al. 2004). For the CP *Ir/iGluR* members, we required the presence of the two ligand-binding domains, namely, PF00060 (ligand-gated ion channel) and PF10613 (ligand ion channel L-glutamate- and glycine-binding site), which are present in all *Ir/iGluR* subfamilies, i.e., kainate, AMPA, NMDA, conserved IRs (Ir25a/Ir8a), and divergent IRs (Croset et al. 2010). The third domain exhibited by some members of the family, PF01094 (ANF receptor), was not used in this step. The remaining sequences that were free of stop codons and did not pass the length filter criteria were classified as incomplete proteins (IP set). Finally, the CP and IP sequences exhibiting some in-frame stop codons (that could represent pseudogenes, among other features; Ψ) were incorporated into two extra data sets (CPΨ and IPΨ sets, respectively).

We used three different estimators of the number of copies of a particular CS family (family size). In addition to the straightforward number of CPs in a particular genome (S_{CP}), we also determined the minimum number of sequences that could be unequivocally attributed to different functional genes (S_{MIN}) and the maximum number of members in cases where all the incomplete protein fragments were actually different functional genes (S_{MAX}). We estimated these numbers by aligning all protein sequences (both CP and IP) within a family using the CS-F-HMM profile as a guide and examining the matching distribution of all fragments aligned along the protein. The S_{MIN} was obtained by adding to the total number of sequences present in the CP set, the minimum number of sequences of the IP set that could be unequivocally attributed to different family members. This minimum amount was determined by counting the number of partial sequences aligned in the most covered protein region of the CS-F-HMM profile-guided MSA. The S_{MAX} is the total number of both CP and IP copies identified (supplementary table S1*B* and *C*, Supplementary Material online).

Phylogenetic Analyses

As the divergence between some members of the same CS family is huge (i.e., their most recent common ancestor traces back far before the split of the major arthropod lineages, ~600 Ma; Hedges et al. 2006), building a reliable MSA to estimate the phylogenetic relationships is not straightforward. To address this long-standing problem, we applied the MSA-free HMM distance-based method (Bogusz and Whelan 2017) implemented in the PaHMM-Tree software, which outperforms MSA-based methods when dealing with the high alignment uncertainty that is usually associated with large divergences. All the phylogenies except those of the IR family (see Results for more details about this family) were based on complete sequences. We used the iTOL web server (Letunic and Bork 2007) to format and display the trees.

Gene Turnover Rates

We estimated the gene family turnover rates using a gene tree–species tree reconciliation approach. The ultrametric species tree required for the analysis was inferred by fitting the amino acid variation of all 88 putative single-copy orthologs to the most accepted topology for the 11 species. For the analysis, we used OrthoMCL (v2.0.9; Li et al. 2003) to identify 1:1 orthologs by clustering the sequences by similarity and then generated an MSA (for each ortholog group) with T-Coffee v11.00 (mcoffe mode; Notredame et al. 2000). After filtering the MSAs with trimAl v1.4 (-automated1 option; Capella-Gutiérrez et al. 2009), we estimated the best-fit amino acid substitution model for each MSA with the program jModelTest based on the Akaike information criteria for model selection (Guindon and Gascuel 2003; Darriba et al. 2012) and concatenated all MSA, keeping the individual coordinate information to be used as a partition for the phylogenetic analysis. We used RAxML software (option –f e) to obtain ML estimates of branch lengths and r8s software v 1.80 (Sanderson 2003) to linearize the unrooted ML using the penalized likelihood algorithm. For the last step, we constrained the ages of two internal nodes according to the fossil calibrations: 1) the root (on the range 528–445 Myr; Dunlop and Selden 2009) and 2) the split between scorpions and spiders (at a minimum of 428 Myr; Jeyaprakash and Hoy 2009).

We analyzed the family turnover rates for the two largest gene families in *Arachnida*, *Gr* and *IrIiGluR*, using a gene tree–species tree reconciliation approach. For each family and lineage, we estimated separately the birth (β) and death (δ) rates, which measure the number of sequence gains and losses per sequence per million years, respectively. For the global analysis, we estimated the average values across all branches, excluding *Li. polyphemus*, which was used to root the tree. We used the software OrthoFinder (Emms and Kelly 2015) to obtain orthogroups (i.e., all groups of N: N orthologs) and gene trees to calculate the number of gene gain and loss events in each lineage with the program Notung (Chen et al. 2000). Finally, we estimated the global turnover rates (β and δ) from these events using formulas 1 and 2 in Almeida et al. (2014), whereas the net turnover rates (Δ) were directly estimated as $\Delta = \beta - \delta$.

Results

The Chemosensory Subgenome of Chelicerates

Our comprehensive search protocol revealed 6,026 CS protein-coding sequences across the 11 surveyed chelicerate genomes (supplementary table S1*A*, Supplementary Material online). Surprisingly, nearly 85% of them (5,086) had previously inaccurate genome annotations, including 4,131 non-annotated sequences (without a GFF record) and another 955 that, despite having structural annotation data in the GFF file, lacked functional information (as putative CS proteins) in the GFF field. Nevertheless, only 2,646 of the 6,026 sequences (supplementary table S1*B*, Supplementary Material online) encoded complete (or nearly complete) CS proteins free of stop codons (CP set). Among the remaining sequences, 1,895 were incomplete (but without stop codons in frame) (IP set) and 1,485 showed one or more premature stop codons (including both CP and IP sequences). Globally, the actual number of putative functional CS genes ranged from 4,255 (S_{MIN}) to 4,541 (S_{MAX}), although only 2,646 of them were complete (S_{CP}) (supplementary table S1C, Supplementary Material online). Remarkably, although canonical insect *Obp* and *Or* genes were absent in chelicerate genomes, we found a huge and unexpected number of novel *Gr*-coding (108 uncharacterized peptides plus 3,331 novel genomic sequences) and *IrIiGluR*-coding (525 plus 694) sequences. Furthermore, it is noteworthy that *Csp* members were absent in all genomes, except in the tick *I. scapularis*, and *Ccp* family members were identified only in spiders and scorpions (fig. 1).

Chemoreceptors

We found that the *Gr* family is the largest CS gene family in chelicerates ($S_{MIN} = 3{,}074$, $S_{MAX} = 3{,}157$, and $S_{CP} = 2{,}032$, considering only putative functional sequences; fig. 1; supplementary table S1*B*, Supplementary Material online). Moreover, we also identified 1,097 putative *Gr* pseudogenes (see Discussion). Remarkably, there are extraordinary differences in the family size across chelicerates; although some species exhibit >400 copies, such as the scorpion *C. exilicauda* ($S_{MIN} = 832$), the tick *T. urticae* ($S_{MIN} = 469$) or the spider *P. tepidariorum* ($S_{MIN} = 643$), others have <60, such as *I. scapularis* ($S_{MIN} = 57$) and *Li. polyphemus* ($S_{MIN} = 58$) (supplementary table S1C, Supplementary Material online). These results cannot be explained by putative differences in the assembly quality across genomes because the same trend was observed with S_{MAX} and S_{CP} values. In fact, there is no relationship between the values of our three estimates of the real number of *Gr* genes across genomes and the N50, the number of scaffolds or the number of predicted peptides in these genomes (supplementary table S1C, Supplementary Material online). Strikingly, even the most closely related species, the spiders *La. hesperus* and *P. tepidariorum*, greatly differ in their repertory size (fig. 1), revealing a highly dynamic evolution. These differences are clearly shown in the phylogenetic tree as large monophyletic groups (mostly species-specific clades). Despite these findings, the tree also reveals a distinctive monophyletic group of apparently less dynamic sequences with representatives from all chelicerates (fig. 3; supplementary fig. S1, Supplementary Material online). However, we did not detect any GR protein closely related to the functionally characterized carbon dioxide, sweet taste, and fructose insect receptors in chelicerates (Jones et al. 2007; Miyamoto et al. 2012; Fujii et al. 2015).

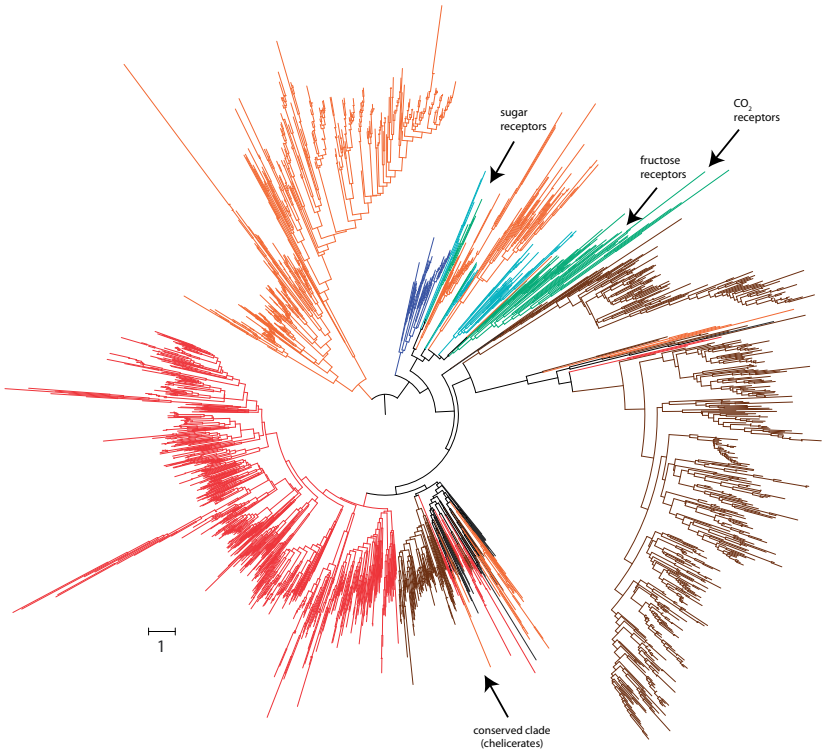


Fig. 3.—Phylogenetic tree of the *Gr* family members across arthropods. The different species are depicted in colors as in figure 1. The scale bar represents one amino acid substitutions per site.

The *Ir/iGluR* is the second largest CS family ($S_{CP} = 323$, $S_{MIN} = 825$, and $S_{MAX} = 979$). Again, but less pronounced than in the *Gr* family, we also detected a highly uneven distribution of copies across lineages. Interestingly, the repertory sizes of these two families do not correlate across chelicerates (Pearson correlation, *P*-value > 0.05); for instance, *T. urticae* encodes very few *Ir/iGluR* copies ($S_{MIN} = 19$) but a large number of *Gr* genes ($S_{MIN} = 469$). Similar to the *Gr* family, the relationship of the *Ir/iGluR* family size across species is very similar regardless of the use of S_{CP}, S_{MIN}, or S_{MAX} values, suggesting that the assembly quality has no influence.

The phylogenetic analysis using sequences with the complete ligand channel domain reproduced the established relationships of the five major arthropod *Ir/iGluR* subfamilies (fig. 4; supplementary fig. S2, Supplementary Material online; Croset et al. 2010; Vizueta et al. 2017). The gene topology allowed us to identify 249 IR proteins (or truthful IR set, t-IR) (200 with the two ligand-binding domains plus another 49 with only the ligand channel domain; supplementary table S1C, Supplementary Material online), which would represent the minimum number of functional IR copy candidates to perform a chemosensory function. The phylogenetic analysis also revealed the absence of members of the *Ir25a/Ir8a*-conserved IR subfamily in *M. martensii*, *S. mimosarum*, *A. geniculata*, and *La. hesperus*. However, a more comprehensive analysis of the IP set revealed that, in fact, all these species encode one IR25a receptor (supplementary table S2 and fig. S3, Supplementary Material online). Interestingly, we failed to detect any putative homologs of IR8a in all chelicerates, except in the horseshoe crab *Li. polyphemus* (LpolIR11 sequence). Still, we could detect putative homologs of two *Drosophila* antennal IRs, IR93a and IR76b. The first member was identified in all species, excluding *A. geniculata* and *S. mimosarum*, whereas IR76b was present in *Daphnia*, the horseshoe crab, the two scorpions and the spiders *P. tepidariorum* and *La. hesperus* (supplementary table S2 and fig. S3, Supplementary Material online). Nonetheless, we did not find putative homologs of the other *Drosophila* antennal IRs with

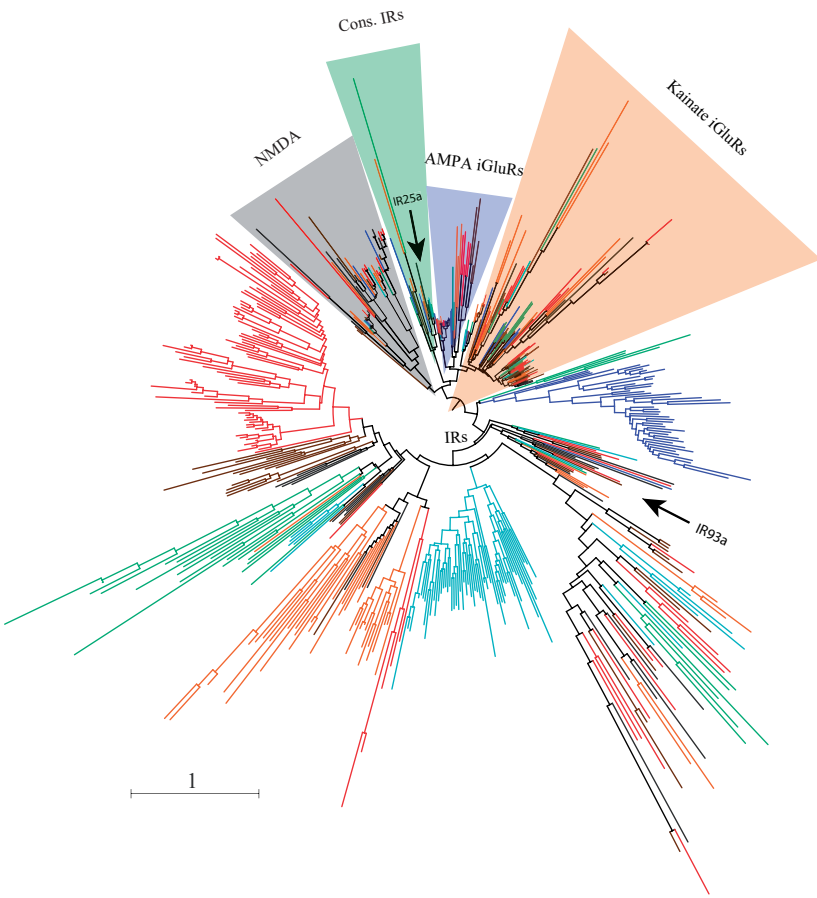


Fig. 4.—Phylogenetic tree of the *IrliGluR* family members across arthropods. The tree is based on LCD domain sequences (PF00060). Different lineages are colored as in figure 1. The three main subfamilies of iGluRs and the conserved IR clade are shaded in different colors. The scale bar represents one amino acid substitution per site.

orthologous copies in insects, such as IR21a and IR40a (Croset et al. 2010; Eyun et al. 2017).

Other Chemosensory Families

We identified several novel and complete OBP-like encoding sequences in chelicerates (fig. 1; supplementary table S1A, Supplementary Material online). In addition to the described members in *I. scapularis*, *M. occidentalis*, *S. mimosarum*, and *S. maritima* (Renthal et al. 2017; Vizueta et al. 2017), we identified a total of 26 new (out of 30) OBP-like proteins in chelicerates. All the chelicerates encode at least one member of this family, with repertory sizes ranging from 1 to 4 copies. Additionally, and very surprisingly, we detected 19 novel (out a total of 21) *Obp-like* genes in the centipede *S. maritima*. Our phylogenetic analysis of canonical OBP (from insects) and OBP-like proteins (fig. 5, supplementary fig. S4, Supplementary Material online) does not support the reciprocal monophyly of these two gene families. Although some OBP-like sequences (such as MoccOPBl2, IscaOBPl2 and PtepOBPl3) are phylogenetically close to the OBP Plus-C subfamily, others, for example, DmelOBP99c (a member of the insect minus-C subfamily), are more related to the chelicerate OBP-like sequences than to the insect OBP sequences.

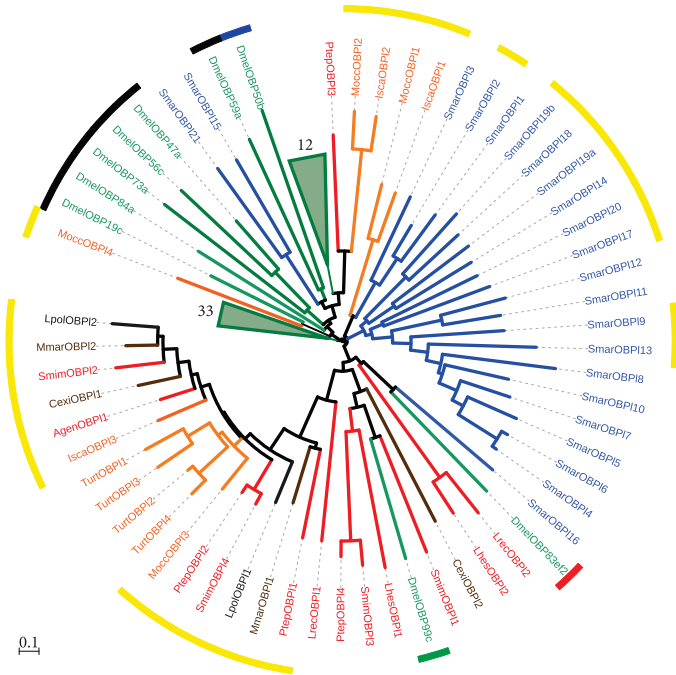


Fig. 5.—Phylogenetic relationships of the *Obp*-like and insect (*D. melanogaster*) *Obp* family members. Lineages and species names are colored as in figure 1. For clarity, two *D. melanogaster* nodes with 12 and 33 descent sequences are collapsed. The color of the inner circle indicates the *Obp* subfamily: Classic (black), Minus-C (green), Plus-C (blue) and Dimer (red). The outer circle in yellow indicates the members from noninsect species with PBP/GOBP domain (IPR006170). The scale bar represents 0.1 amino acid substitutions per site.

Moreover, the phylogenetic analysis revealed three major clades, each almost exclusively containing sequences of the given arthropod subphylum (i.e., *D. melanogaster*, *S. maritima*, and chelicerates).

The size of the *Npc2* family has remained relatively constant during the diversification of the major chelicerate lineages, ranging from 10 to 20 (S_{MIN} values, supplementary table S1C, Supplementary Material online), with the outstanding exception of *T. urticae*, which encodes 47 genes. Nevertheless, nearly half of the *Npc2* members of some species are incomplete fragments or show premature stop codons, resulting in much greater difficulty in drawing a firm conclusion about the real sizes of this family compared with the other families. In this case, we found a strong positive correlation between N50 and S_{CP}, S_{MIN}, and S_{MAX} values (Pearson correlation coefficient, $r > 0.80$; $P < 0.05$; supplementary table S1C, Supplementary Material online), indicating that the observed variation in the number of *Npc2* genes across species is clearly associated with genome assembly continuity. This result is probably due to the fact that the length of the genomic region that includes the target sequences of the similarity searches is the longest (jointly with the *Cd36-Snmp* family, see below) among the families surveyed in this work. Unlike chemoreceptors and *Obp*-like members, NPC2 proteins are not arranged in large species-specific phylogenetic clades (supplementary fig. S5, Supplementary Material online), suggesting a less dynamic evolution of this family compared with chemoreceptors and OBP-like proteins.

Our searches for members of the recently discovered *Ccp* gene family (Vizueta et al. 2017) only provided positive results in spiders and in *Centruroides exilicauda* (the Bark scorpion), although the sequence identity of the copy detected in this last species is low. We found important differences in family size across species, from 2 in *Lo. reclusa* to 21 in *P. tepidariorum* (S_{MIN}). Like in *D. silvatica*, most CCPs exhibited an identifiable signal peptide sequence and a conserved cysteine pattern, supporting their putative role in the extracellular binding and transport of chemical cues (Vizueta et al. 2017).

The phylogenetic analysis of this family revealed relatively short branches and clades likely representing orthologous genes (supplementary fig. S6, Supplementary Material online). Even so, the 21 copies of *P. tepidariorum* (11 of them forming a species-specific clade) is a remarkable exception and could be associated with an adaptive event linked to this family in this lineage. The high-quality assembly and annotation of the *P. tepidariorum* genome may be good enough to have a closer look at the genomic location of *Cpp* genes and to search in this family for signatures of the lineage-specific bursts of tandem duplications stated by Schwager et al. (2017).

The *Cd36-Snmp* Family

The *Cd36-Snmp* family size has also remained relatively constant during the diversification of chelicerates, especially with respect to the S_{MAX} values (ranging from 8 to 19). Nevertheless, as in the *Npc2* family, nearly half of the positive hits encode incomplete proteins, most of which are in spiders and scorpions (supplementary table S1*B*, Supplementary Material online). Consistent with the large size of the target genomic regions of this family, we also found a positive correlation between N50 and S_{CP} and S_{MIN} (but not S_{MAX}) values for this family (Pearson correlation coefficient, $r > 0.56$; $P < 0.05$; supplementary table S1*C*, Supplementary Material online), although weaker than in the case of NPC2. The phylogenetic analysis (supplementary fig. S7, Supplementary Material online) showed that only one of three phylogenetic clades described by Nichols and Vogt (2008) has remained monophyletic across all arthropods (i.e., the group including the SNMP protein of *D. melanogaster*). However, many sequences do not form monophyletic groups and, therefore, cannot be unambiguously assigned to a given subfamily group, suggesting a more complex grouping than those observed in insects (Nichols and Vogt 2008).

Gene Turnover Rates of Chemoreceptors

We estimated gene family turnover rates for the two largest *Chelicerata* gene families, *Gr* and *Ir/iGluR*, using *Li. polyphemus* to root the tree (fig. 6, supplementary fig. S8, Supplementary Material online). As the analysis could have been compromised by the use of three different estimates of family size (per CS family), we first evaluated the behavior of these size estimates with respect to the turnover rates. We found that the number of gene duplications and losses calculated using S_{CP} (only for the *Gr* family), S_{MIN}, and S_{MAX} values strongly correlated across lineages ($r > 0.94$; *P*-values $< 10^{-5}$); therefore, we did not expect important relative rate differences among the three estimates. Thus, we calculated birth and death rates only with S_{MIN} because this estimate likely represented the true number of copies in most genomes.

We found that the global (across all phylogenetic tree) gene turnover rates of *Gr* and *Ir/iGluR* showed important differences (supplementary fig. S8, Supplementary Material online). In *Gr*, the net turnover rates were positive ($\Delta = 0.003$), indicating an overall expansion of gustatory receptor repertory during arachnid diversification. In contrast, the *Ir/iGluR* family showed an overall contraction ($\Delta = -0.002$). These results should be considered with caution because global turnover rates are strongly affected by the presence of specific phylogenetic branches with extreme values. In the *Gr* family, for instance, the external lineages leading to *T. urticae* ($\beta = 0.015$), *C. exilicauda* ($\beta = 0.030$), and *P. tepidariorum* ($\beta = 0.030$) have β values that are much higher than the global rates ($\beta = 0.007$); in contrast, other branches, such as the internal lineage leading to acari ($\delta = 0.008$) and the external lineage leading to *La. hesperus* ($\delta = 0.007$), show death rates that clearly exceed global estimates ($\delta = 0.004$).

The *Ir/iGluR* family exhibits smaller turnover rate differences among the lineages than those observed for *Gr*. Even so, the external branches of *C. exilicauda* ($\beta = 0.005$), and especially of *P. tepidariorum* ($\beta = 0.011$), are clear outliers and the only ones that show a clear expansion of the *Ir/iGluR* repertory during the diversification of arachnids. It should be noted that the *Ir/iGluR* data set includes the sequences of five subfamilies of this highly functional, diverse family of receptors, which show very dissimilar turnover rates in insects. In fact, the *Ir* subfamily, which is the only subfamily encoding putative chemosensory receptors, is the most dynamic family of insects. Therefore, to disentangle subfamily-specific effects, we estimated the gene turnover rates using only the IR copies from S_{MIN} and the t-IR set (fig. 4). As expected, birth and death rates estimated from the S_{MIN} and t-IR sets did not show big differences (results not shown), suggesting a major effect of the *Ir* subfamily on gene turnover estimates in the *Ir*/iGluR family. Indeed, the t-IR estimates were even more variable across lineages than those obtained for the whole family, especially for birth rates, with slightly higher average rates. Especially noteworthy is the case of the *P. tepidariorum* lineage, which not only confirmed the findings of the S_{MIN} set analysis but also showed that the gene number expansion (supplementary fig. S2, Supplementary Material online) was definitively caused by the birth of new *Ir* genes (t-*Ir* set based estimates, $\beta = 0.020$, $\delta = 4 \times 10^{-4}$).

Discussion

The early diversification of arthropods predated the colonization of land by animals (Rota-Stabelli et al. 2013). Chemical communication strategies associated with this terrestrialization, therefore, should have been invented several times independently in their major lineages (*Hexapoda*, *Crustacea*, *Myriapoda*, and *Chelicerata*). It is likely that proteins involved in the first peripheral chemosensory perception steps, which are commonly associated with medium-size gene families, played a central role. Hence, these gene families represent an important fraction of arthropod genomes and contribute

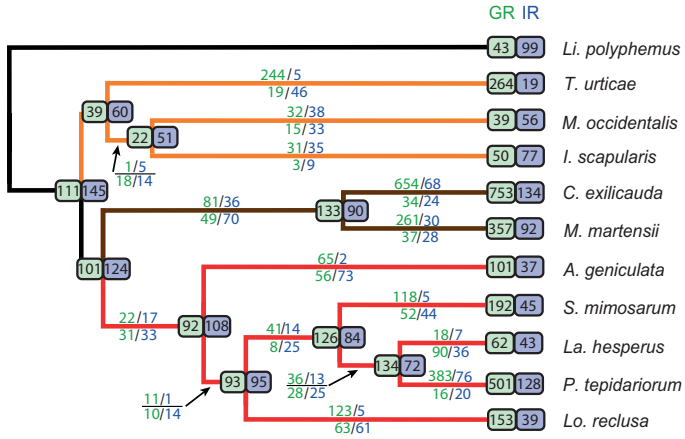


Fig. 6.—Gene turnover of chemoreceptors across chelicerates. Estimates obtained from the data set used to estimate S_{MIN}. Numbers above and below each branch indicate lineage-specific gene duplications and losses, respectively. Green, GR family; blue, IR/iGluR family. Estimates in very short and outgroup branches have large uncertainty and are not showed. Numbers in the ancestral nodes show the estimated family sizes. Numbers at the tips indicate the number of sequences used for the analysis; such values can differ from S_{MIN} because only sequences that clustered in an orthogroup (with three or more sequences) were included in the analysis.

significantly to gene turnover dynamics in insects (Sánchez-Gracia et al. 2009, 2011). The recent availability of the complete genome sequences from various chelicerates has provided insights into their CS family members. Nevertheless, the quality of the genome assembly and functional annotation is far from satisfactory. Some genomes are highly fragmented, with an absence of functional annotations or annotations obtained using only nonexhaustive automated protocols. Here, we report the first comparative analysis of the actual copy number and gene turnover evolution of CS families in 11 nonhexapod genomes. This study is in fact the first comprehensive comparative genomics study that, although enriched in *Arachnida* species, covers most of the chelicerate diversity (see Eyun et al. [2017], Palmer and Jiggins [2015], and Sanggaard et al. [2014] for examples of previous studies based on many fewer genomes).

The Outstanding Chemoreceptor Repertory of *Chelicerata* Genomes

The most important challenge for understanding gene family evolution is having well-characterized copies and accurate functional annotations of their members. This is particularly relevant when using highly fragmented genome assemblies generated from short-read sequencing data. To circumvent this problem, we applied a very comprehensive identification and characterization protocol that combined both protein and DNA sequence data, including HMM profiles and protein domain signatures, in a series of sequential searches with accurate filters based on our biological knowledge of the CS system. Our study revealed a surprisingly large number of novel *Gr*- and *Ir*-encoding sequences. This feature can be mostly explained by the poor functional annotation status of some genomes. In fact, in those genomes in which CS families had been explicitly characterized (the three acari species, *D. melanogaster*, *D. pulex*, and *S. marítima*), our search protocol largely matched with previously annotations. This characteristic, therefore, indicated that the novel CS-encoding sequences were not false positives caused by a misleading search protocol.

We also found that some of the newly identified CS genes were highly fragmented, which is also a consequence of the low quality of assemblies and, therefore, of the poor annotation of gene structures in most surveyed genomes. Most genes are distributed across many different scaffolds, preventing the calculation of the exact number of functional copies in a particular genome. This feature led us to define three repertory size statistics, which not only provided an approximate idea of true values but also allowed for harmonized comparisons across genomes and lineages. As expected, the largest discrepancy occurred between size estimates based on complete genes (S_{CP}) and those including information of incomplete gene fragments (S_{MIN} and S_{MAX}). Despite this difference, however, all three data sets yielded very similar estimates of gene turnover rates; therefore, all of them are good approximations of true CS family sizes and are appropriated to study gene family dynamics across chelicerates. Although S_{MIN} and S_{MAX} values were generally similar, two families showed very

important discrepancies: *Ir/iGluR* and *Cd36-Snmp*. These discrepancies could be explained by the fact that these genes (and the encoding region including introns) are larger than in the other families, and therefore, it is more likely that the encoding region was fragmented in different scaffolds. In fact, this effect was not observed in genomes with more contiguity (based on the N50 values of the genome assemblies), as observed in *T. urticae*, *M. occidentalis*, *S. mimosarum*, and *P. tepidariorum*. Finally, we also found numerous sequences with in-frame stop codons, which we have preliminarily classified as putative pseudogenes. It should be taken into account that not all sequences with evidence of stop codons must be nonfunctional copies; indeed, some of these stop codons may be introduced during gene assembly from dispersed TBlastN hits (which has been done in a semiautomatic way). Only with the use of additional, high-quality assembled genomes will it be possible to obtain accurate information concerning the nature and number of these putative pseudogenes.

CS Gene Turnover in Chelicerates: Complex Evolutionary Dynamics

We have shown that although chelicerate have larger *Gr* gene repertories than nonchelicerates, the estimated birth and death rates for the *Gr* family are almost the same as those in insects (Almeida et al. 2014). The disparate family sizes might be explained by former differences in the ancestors of each of these two lineages. In fact, at least two ancient and independent whole-genome duplications (WGD) have been proposed for chelicerates, one in the ancestor of spiders and scorpions (~450 Ma; Schwager et al. 2017), and the other likely occurred in the lineage of horseshoe crabs (Kenny et al. 2016; Schwager et al. 2017). Thus, it is tempting to hypothesize that evolutionary forces and genomic mechanisms underlying the long-term birth and death dynamics of chemosensory families were essentially the same in all arthropods, although eventually promoted by lineage-specific genome-scale events such as WGD. Nevertheless, not all of our results are compatible with such an evolutionary scenario. For instance, the results obtained for the *Ir* subfamily do not agree with those observed for *Gr*. The birth and death rates of these putative chemoreceptors differ between chelicerates and nonchelicerates, and they do not show the footprint of the WGD preceding the diversification of spiders and scorpions. In fact, net turnover rate of this family has the opposite pattern as GRs, suggesting an important contraction of ionotropic receptors in chelicerates.

Furthermore, the occurrence of WGD events could not satisfactorily explain the full evolutionary history of most of the surveyed families, not even for the *Gr* family. For instance, *T. urticae* shows very high GR repertoires ($S_{MIN} = 469$) and a very low IR ($S_{MIN} = 6$) compared with the other acari, and this pattern is unequivocally not explained by the use of a particular family size S_{MIN} statistic (the three estimators point to the same feature). Although we cannot completely rule out the possibility of a WGD in this lineage, there is no compiled evidence in support of this phenomenon (Grbić et al. 2011; Kenny et al. 2016). Second, the closest phylogenetic lineages in our study (*La. hesperus* and *P. tepidariorum*, with the most recent common ancestor tracing back approximately 100 Ma) show enormous differences in *Gr* and *Ccp* family sizes. Finally, estimation of the turnover rates in a pair of phylogenetically close species (*C. exilicauda* and *M. martensii; La. hesperus* and *P. tepidariorum*) is difficult to reconcile with a constant birth and death process. Therefore, the evolutionary process was rather complex and cannot be entirely explained by WGD. Here, we have demonstrated that other processes affecting specifically chemosensory families, such as long-term birth-and-death evolution associated with high turnover rates occurred in parallel to these whole genomic changes. In addition, more episodic, and probably lineage-specific, expansions and/or contractions also contributed to determine current sizes, as suggested in other studies (Chipman et al. 2014; Schwager et al. 2017). In order to know the relative role of these different processes in shaping actual CS family sizes and their functional meaning, it is imperative to improve the quality of existing genomes and include in the analysis new, more closely related genomes (i.e., increase the phylogenetic coverage).

Phylogenetic Analysis of CS Genes in Arthropods

Despite the above-mentioned limitations, our phylogenetic analysis can shed light on the diversification pattern of CS families. As arthropod CS families are very old and many of their members, especially chemoreceptors, are distantly related, the use of the standard MSA alignment method could be inappropriate for building robust phylogenies. A common method to circumvent this problem is filtering poorly aligned positions and, therefore, considering only highly conserved sites for phylogenetic analyses (Croset et al. 2010; Wu et al. 2016). This approach nevertheless results in a significant loss of relevant amino acid positions that likely contain valuable information on functional and structural features related to the molecular specificity and diversification. Here, we used, for the first time in highly divergent CS families, a method to estimate gene trees using an MSA-free approach, which takes into account alignment uncertainty. For the sake of comparison, we reconstructed the same phylogenetic trees using RAxML based on HMM profile-guided MSAs (Stamatakis 2014: Supplementary file 4). Major differences between PaHMM-Tree and RAxML were found at internal nodes and nodes with low bootstrap support in ML trees (< 70% from 500 replicates). Although bootstrap values increased when filtering poorly aligned positions (Capella-Gutiérrez et al. 2009), the number of informative sites retained after removing these unreliable positions was very low, causing the ML

trees to be based on a very small number of positions. These trees may not be reflecting the real evolutionary history of the chemosensory proteins. Besides, for very large families, such as the *Gr*, the bootstrap analysis was unfeasible in the practice due to excessive computation times. Given that PaHMM-Tree is an alignment-free approach, which allow us to utilize all the amino acid positions to reconstruct the trees, and that the results obtained by Bogusz and Whelan (2017) point to a better performance of this approach for highly divergent sequences without the need for a previous filtering step, here, we decided to report the results based on this method. However, a more exhaustive study comparing these and other tree reconstruction methods, using both real and simulated data and under different degrees of divergence, would be necessary to know whether this method actually improves the phylogenetic analysis. Our phylogenetic analysis correctly recovered all previously known (and accepted) relationships among subfamilies and revealed new aspects of the diversification of CS genes.

We found that chelicerates virtually have their own GR repertoires with almost no phylogenetic clade containing members of insects, crustaceans, and myriapods. In fact, we did not find homologs of any of the GR functionally characterized in insects. Apparently, chelicerate genomes do not encode any protein sequence close to *Drosophila* sugar, fructose, or carbon dioxide receptors (Jones et al. 2007; Miyamoto et al. 2012; Fujii et al. 2015), questioning their ability to detect these substances. Nevertheless, chelicerates might be using other phylogenetically distant gustatory receptors to perform these tasks. Yet, the presence of a monophyletic clade with more conserved GR chelicerate sequences would suggest the existence of some other important biological function played by these receptors. The members of this clade could have a highly relevant function in chelicerates, evolving under lower evolutionary rates despite the tremendous diversification of this subphylum. Future functional studies combined with new evidence based on greater coverage phylogenetic analysis will definitely shed light on this interesting hypothesis.

Another remarkable result is the verification that most GR receptors found in species with very large repertories such as in *P. tepidariorum* or in *C. exilicauda* are monophyletic, pointing to important bursts of gene duplication events in relatively recent time periods. These events probably represent adaptive expansions of the gustatory repertory associated with chemosensory diversifications. In other cases, such as in *T. urticae* lineage, apparent species-specific family expansions might be just an artefact caused by the continued effect of the birth-and-death process in a very long terminal branch (i.e., reflecting the low phylogenetic coverage of this part of the tree).

Although the general phylogenetic pattern observed in the IR is very similar to that of the GR, we detected some *Ir* members with relatively conserved sequences across all arthropods. We can hypothesize that these receptors should have a very relevant and not easily replaceable function. For instance, IR25a, a receptor found in all arthropods surveyed to date, is a broadly expressed protein involved in trafficking to the membrane of other IRs in olfactory and taste organs that has been proposed to have also a coreceptor function in the membrane (Joseph and Carlson 2015). We also found a putative ortholog of IR8a in the horseshoe crab *Li. polyphemus*, which led us to reformulate the hypothesis of Eyun et al. (2017) suggesting that this member arose in the ancestor of myriapods and pancrustaceans, tracing back its origin, again, to at least the ancestor of arthropods.

Our analysis also supports the presence of a group of IR76b homologs outside the insect clade (Eyun et al. 2017) which was likely present in the arthropod ancestor. This receptor, proposed to play a coreceptor function for other IRs and associated with a gustatory function as a detector of low salt concentrations (Zhang et al. 2013), has been identified in all chelicerates except in the acari and some spider clades. Its absence in these arthropod groups suggests a secondary loss in the ancestor of these lineages. However, we could not fully refute the possibility that we were unable to detect this member in these genomes, especially in spiders, because of assembly fragmentation. Our current phylogenetic analysis failed to detect putative homologs of IR21a and IR40a in chelicerates. Though we found some week evidence for homologs of these receptors in the transcriptome of the spider *D. silvatica* (Vizueta et al. 2017), we rely more in the analysis applied herein, which is most comprehensive and uses an alignment-free method based on HMM profiles to generate the trees. These new evidences, together with previous genomic analyses, would indicate the presence of IR21a exclusively in panarthropods (Eyun et al. [2017] have recently found a putative homolog of the IR21a protein in copepods) and of IR40a exclusively in insects.

Notably, our study shows that all chelicerates and the centipede *S. maritima* carry members of the *Obp*-like family, a gene family that is closely related to insect OBPs (Renthal et al. 2017; Vizueta et al. 2017). This family, which is absent in crustaceans, might represent a remote homolog of canonical insect OBPs. The close relationship of a *Drosophila* minus-C OBP within an OBP-like chelicerates clade, in agreement with the results of Renthal et al. (2017) based on the disulfide bonding pattern, suggests that this subfamily represents an ancestral state of an OBP. Nonetheless, we cannot completely ignore the possibility that the similar sequence arose by structural convergence. As a canonical OBP, OBP-like has a signal peptide region, a predicted globular protein with the characteristic cysteine patterns of OBPs, and predicted folding similar to that of insect OBPs. Moreover, some experimental results have also confirmed the expression of some *Obp*-like members in specific chelicerates chemosensory appendages (Renthal et al. 2017). All compiled evidence, therefore, suggests that chelicerates and

myriapod OBP-like may have a similar function to canonical OBPs, such as in solubilizing and transporting chemical cues. Regardless, the extraordinarily large repertory observed in *S. maritima* clearly merits further investigation. This is especially interesting because the genome paper of *S. marticima* reported a high number of tandem duplications (Chipman et al. 2014).

Intriguingly, we did not find CSP-encoding genes in the surveyed chelicerates, except the single copy found in the tick *I. scapularis* (Gulia-Nuss et al. 2016). Although Eyun et al. (2017) reported some sequences encoding CSP proteins in the bark scorpion *C. exilicauda* and the spider *La. hesperus*, our analysis of such sequences could not unequivocally establish that they encode real CSP proteins; indeed, these sequences are very short with multiple in frame stop codons and do not exhibit the characteristic cysteine CSP pattern, suggesting a false positive result. Our analysis also allowed us to identify members of the *Ccp* gene family in spiders, as well as a remote homolog in the bark scorpion *C. exilicauda*, suggesting that the origin of this rapidly evolving gene family traces back to the ancestor of these two groups. Remarkably, we observed a large expansion of some members (a lineage-specific expansion) in the house spider *P. tepidariorum*, a feature that reflects its greater number of chemoreceptors. We have established that the CCP-encoding genes have a signal peptide fragment and similar folding characteristics to the insect OBP and are differentially expressed in the putative chemosensory appendices of the spider *D. silvatica* (Vizueta et al. 2017). Therefore, although their actual function is unknown, it is tempting to assign a putative function to the transport and solubilization of chemical cues, a functional role equivalent to that of the canonical OBP, Nevertheless, given that the *Ccp* is a rapidly evolving gene family that emerged in some derived chelicerate lineages, it could provide new insights into the extracellular-binding protein functions and their roles in diversification and adaptation in arthropods.

Conclusions

Noninsect arthropods comprise a significant portion of earth's biodiversity and include many species of economic and medical importance. Here, we conducted the first comprehensive comparative genomic analysis across 11 genomes of this old lineage and the first of this magnitude outside of insects. Despite that the high fragmentation of genome drafts prevented us from establishing the exact number of chemosensory genes in each species, our exhaustive search protocol exposed an unprecedented huge number of new family members. Remarkably, many of these new genes were not characterized or even not detected before and most of them encode chemoreceptors. Moreover, we found a remarkable disparity in chemoreceptor repertories across species that is difficult to explain without invoking lineage-specific adaptive expansions probably related with sensory diversification processes. Characterizing the intragenomic dynamics and the specific function of these recently expanded chemosensory genes is an exciting prospect that jointly with the improvement of existing genome assemblies and the reduction of the phylogenetic gap will allow researchers to move forward in the knowledge of chelicerate genomics and biology. This work aims to contribute to this advance and hopes to be the starting signal for many future comprehensive comparative genomic studies in a group of animals as fascinating as unknown.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by the Ministerio de Economía y Competitividad of Spain (CGL2013-45211 and CGL2016-75255) and the Comissió Interdepartamental de Recerca I Innovació Tecnològica of Catalonia, Spain (2014SGR-1055). J.V. was supported by an FPI grant (Ministerio de Economía y Competitividad of Spain, BES-2014-068437) and J.R was partially supported by ICREA Academia (Generalitat de Catalunya). The authors declare that they have no competing interests.

Author Contributions

A.S.-G. and J.R. conceived and designed the study. J.V. analyzed the data. J.V., J.R. and A.S.-G. wrote the manuscript.

Literature Cited

Adams MD, et al. 2000. The genome sequence of *Drosophila melanogaster*. Science 287:2185–2195.

Almeida FC, Sánchez-Gracia A, Campos JL, Rozas J. 2014. Family size evolution in *Drosophila* chemosensory gene families: a comparative analysis with a critical appraisal of methods. Genome Biol Evol. 6(7):1669–1682.

Altschul S. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25(17):3389–3402.

Babb PL, et al. 2017. The *Nephila clavipes* genome highlights the diversity of spider silk genes and their complex expression. Nat Genet. 49(6):895–903.

Benton R, Vannice KS, Gomez-Diaz C, Vosshall LB. 2009. Variant ionotropic glutamate receptors as chemosensory receptors in *Drosophila*. Cell 136(1):149–162.

Benton R, Vannice KS, Vosshall LB. 2007. An essential role for a CD36-related receptor in pheromone detection in *Drosophila*. Nature 450(7167):289–293.

Berna AZ, Anderson AR, Trowell SC. 2009. Bio-benchmarking of electronic nose sensors. PLoS One 4(7):e6406.

Bogusz M, Whelan S. 2017. Phylogenetic tree estimation with and without alignment: new distance methods and benchmarking. Syst Biol. 66(2):218–231.

Cao Z, et al. 2013. The genome of *Mesobuthus martensii* reveals a unique adaptation model of arthropods. Nat Commun. 4:2602.
Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25(15):1972–1973.
Chen K, Durand D, Farach-Colton M. 2000. NOTUNG: a program for dating gene duplications and optimizing gene family trees. J Comput Biol. 7(3–4):429–447.
Chipman AD, et al. 2014. The first myriapod genome sequence reveals conservative arthropod gene content and genome organisation in the centipede *Strigamia maritima*. PLoS Biol. 12(11):e1002005.
Colbourne JK, et al. 2011. The ecoresponsive genome of *Daphnia pulex*. Science 331(6017):555–561.
Corey EA, Bobkov Y, Ukhanov K, Ache BW. 2013. Ionotropic crustacean olfactory receptors. PLoS One 8(4):e60551.
Croset V, et al. 2010. Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. PLoS Genet. 6(8):e1001064.
Dani FR, et al. 2011. Odorant-binding proteins and chemosensory proteins in pheromone detection and release in the silkmoth *Bombyx mori*. Chem Senses. 36(4):335–344.
Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 9(8):772.
de Bruyne M, Baker TC. 2008. Odor detection in insects: volatile codes. J Chem Ecol. 34(7):882–897.
Dessimoz C, et al. 2011. Comparative genomics approach to detecting split-coding regions in a low-coverage genome: lessons from the chimaera *Callorhinchus milii* (Holocephali, Chondrichthyes). Brief Bioinform. 12(5):474–484.
Dippel S, et al. 2014. Tissue-specific transcriptomics, chromosomal localization, and phylogeny of chemosensory and odorant binding proteins from the red flour beetle *Tribolium castaneum* reveal subgroup specificities for olfaction or more general functions. BMC Genomics 15(1):1141.
Dunlop JA, Selden PA. 2009. Calibrating the chelicerate clock: a paleontological reply to Jeyaprakash and Hoy. Exp Appl Acarol. 48(3):183–197.
Eddy SR. 2011. Accelerated profile HMM searches. PLoS Comput Biol. 7(10):e1002195.
Eliash N, et al. 2017. Chemosensing of honeybee parasite, *Varroa destructor*: transcriptomic analysis. Sci Rep. 7(1):13091.
Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16(1):157.
Eyun S, et al. 2017. Evolutionary history of chemosensory-related gene families across the Arthropoda. Mol Biol Evol. 34(8):1838–1862.
Forêt S, Wanner KW, Maleszka R. 2007. Chemosensory proteins in the honey bee: insights from the annotated genome, comparative analyses and expressional profiling. Insect Biochem Mol Biol. 37(1):19–28.
Frías-López C, et al. 2015. Comparative analysis of tissue-specific transcriptomes in the funnel-web spider *Macrothele calpeiana* (Araneae, Hexathelidae). PeerJ 3:e1064.
Fujii S, et al. 2015. *Drosophila* sugar receptors in sweet taste perception, olfaction, and internal nutrient sensing. Curr Biol. 25(5):621–627.
Gendreau KL, et al. 2017. House spider genome uncovers evolutionary shifts in the diversity and expression of black widow venom proteins associated with extreme toxicity. BMC Genomics 18(1):178.
Gomez-Diaz C, et al. 2016. A CD36 ectodomain mediates insect pheromone detection via a putative tunnelling mechanism. Nat Commun. 7:11866.
Grbić M, et al. 2011. The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. Nature 479(7374):487–492.
Groh-Lunow KC, Getahun MN, Grosse-Wilde E, Hansson BS. 2015. Expression of ionotropic receptors in terrestrial hermit crab's olfactory sensory neurons. Front Cell Neurosci. 8:1–12.
Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52(5):696–704.
Gulia-Nuss M, et al. 2016. Genomic insights into the *Ixodes scapularis* tick vector of Lyme disease. Nat Commun. 7:10507.
He Q, et al. 2013. The venom gland transcriptome of *Latrodectus tredecimguttatus* revealed by deep sequencing and cDNA library analysis. PLoS One 8(11):e81357.
Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22(23):2971–2972.
Hoy MA, et al. 2016. Genome sequencing of the phytoseiid predatory mite *Metaseiulus occidentalis* reveals completely atomized hox genes and superdynamic intron evolution. Genome Biol Evol. 8(6):1762–1775.
Iovinella I, et al. 2011. Differential expression of odorant-binding proteins in the mandibular glands of the honey bee according to caste and age. J Proteome Res. 10(8):3439–3449.
Ishida Y, Ishibashi J, Leal WS. 2013. Fatty acid solubilizer from the oral disk of the blowfly. PLoS One 8(1):e51779.
Jeyaprakash A, Hoy MA. 2009. First divergence time estimate of spiders, scorpions, mites and ticks (subphylum: Chelicerata) inferred from mitochondrial phylogeny. Exp Appl Acarol. 47(1):1–18.
Jones P, et al. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30(9):1236–1240.
Jones WD, Cayirlioglu P, Kadow IG, Vosshall LB. 2007. Two chemosensory receptors together mediate carbon dioxide detection in *Drosophila*. Nature 445(7123):86–90.
Joseph RM, Carlson JR. 2015. *Drosophila* chemoreceptors: a molecular interface between the chemical world and the brain. Trends Genet. 31(12):683–695.
Käll L, Krogh A, Sonnhammer ELL. 2004. A combined transmembrane topology and signal peptide prediction method. J Mol Biol. 338(5):1027–1036.
Kaupp UB. 2010. Olfactory signalling in vertebrates and insects: differences and commonalities. Nat Rev Neurosci. 11(3):188.
Kenny NJ, et al. 2015. Genome of the rusty millipede, *Trigoniulus corallinus*, illuminates diplopod, myriapod and arthropod evolution. Genome Biol Evol. 7(5):1280–1295.
Kenny NJ, et al. 2016. Ancestral whole-genome duplication in the marine chelicerate horseshoe crabs. Heredity (Edinb) 116(2):190–199.
Kitabayashi AN, Arai T, Kubo T, Natori S. 1998. Molecular cloning of cDNA for p10, a novel protein that increases in the regenerating legs of *Periplaneta americana* (American cockroach). Insect Biochem Mol Biol. 28:785–790.
Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 305(3):567–580.
Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23(1):127–128.
Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13(9):2178–2189.
Li S, et al. 2008. Multiple functions of an odorant-binding protein in the mosquito *Aedes aegypti*. Biochem Biophys Res Commun. 372(3):464–468.
Mille BG, Peigneur S, Predel R, Tytgat J. 2015. Trancriptomic approach reveals the molecular diversity of *Hottentotta conspersus* (Buthidae) venom. Toxicon 99:73–79.
Missbach C, et al. 2014. Evolution of insect olfactory receptors. Elife 3:e02115.

Missbach C, Vogel H, Hansson BS, Groβe-Wilde E. 2015. Identification of odorant binding proteins and chemosensory proteins in antennal transcriptomes of the jumping bristletail *Lepismachilis y-signata* and the firebrat *Thermobia domestica*: evidence for an independent OBP-OR origin. Chem Senses. 40(9):615–626.

Miyamoto T, Slone J, Song X, Amrein H. 2012. A fructose receptor functions as a nutrient sensor in the *Drosophila* brain. Cell 151(5):1113–1125.

Ngoc PCT, et al. 2016. Complex evolutionary dynamics of massively expanded chemosensory receptor families in an extreme generalist chelicerate herbivore. Genome Biol Evol. 8(11):3323–3339.

Nichols Z, Vogt RG. 2008. The SNMP/CD36 gene family in Diptera, Hymenoptera and Coleoptera: *Drosophila melanogaster*, *D. pseudoobscura*, *Anopheles gambiae*, *Aedes aegypti*, *Apis mellifera*, and *Tribolium castaneum*. Insect Biochem Mol Biol. 38(4):398–415.

Nossa CW, et al. 2014. Joint assembly and genetic mapping of the Atlantic horseshoe crab genome reveals ancient whole genome duplication. Gigascience 3(1):9.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol. 302(1):205–217.

Palmer WJ, Jiggins FM. 2015. Comparative genomics reveals the origins and diversity of arthropod immune systems. Mol Biol Evol. 32(8):2111–2129.

Pelosi P. 1996. Perireceptor events in olfaction. J Neurobiol. 30(1):3–19.

Pelosi P, Iovinella I, Felicioli A, Dani FR. 2014. Soluble proteins of chemical communication: an overview across arthropods. Front Physiol. 5:320.

Pelosi P, Zhou J-J, Ban LP, Calvello M. 2006. Soluble proteins in insect chemical communication. Cell Mol Life Sci. 63(14):1658–1676.

Peñalva-Arana DC, Lynch M, Robertson HM. 2009. The chemoreceptor genes of the waterflea *Daphnia pulex*: many Grs but no Ors. BMC Evol Biol. 9(1):79.

Pennisi E. 2017. Spider genes put a new spin on arachnid's potent venoms, stunning silks, and surprising history. Posted in: Biology, Plants and Animals. ; doi:10.1126/science. aar2331.

Qu S-X, Ma L, Li H-P, Song J-D, Hong X-Y. 2016. Chemosensory proteins involved in host recognition in the stored-food mite *Tyrophagus putrescentiae*. Pest Manag Sci. 72(8):1508–1516.

Renthal R, et al. 2017. The chemosensory appendage proteome of *Amblyomma americanum* (Acari: Ixodidae) reveals putative odorant-binding and other chemoreception-related proteins. Insect Sci. 24(5):730–742.

Robertson HM, Wanner KW. 2006. The chemoreceptor superfamily in the honey bee, *Apis mellifera*: expansion of the odorant, but not gustatory, receptor family. Genome Res. 16(11):1395–1403.

Robinson GE, et al. 2011. Creating a buzz about insect genomes. Science 331(6023):1386.

Rota-Stabelli O, Daley AC, Pisani D. 2013. Molecular timetrees reveal a Cambrian colonization of land and a new scenario for ecdysozoan evolution. Curr Biol. 23(5):392–398.

Sánchez-Gracia A, Vieira FG, Almeida FC, Rozas J. 2011. Comparative Genomics of the Major Chemosensory Gene Families in Arthropods. In: *Encyclopedia of Life Sciences (ELS)*. Chichester: John Wiley & Sons, Ltd.

Sánchez-Gracia A, Vieira FG, Rozas J. 2009. Molecular evolution of the major chemosensory gene families in insects. Heredity (Edinb) 103(3):208–216.

Sanderson MJ. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics 19(2):301–302.

Sanggaard KW, et al. 2014. Spider genomes provide insight into composition and evolution of venom and silk. Nat Commun. 5:3765.

Sato K, Touhara K. 2009. Insect olfaction: receptors, signal transduction, and behavior. Results Probl Cell Differ. 47:121–138.

Schwager EE, et al. 2017. The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. BMC Biol. 15(1):62.

Shanbhag SR, et al. 2001. Expression mosaic of odorant-binding proteins in *Drosophila* olfactory organs. Microsc Res Tech. 55(5):297–306.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313.

Vieira FG, Rozas J. 2011. Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: origin and evolutionary history of the chemosensory system. Genome Biol Evol. 3:476–490.

Vizueta J, et al. 2017. Evolution of chemosensory gene families in arthropods: insight from the first inclusive comparative transcriptome analysis across spider appendages. Genome Biol Evol. 9(1):178–196.

Vogt RG, et al. 2009. The insect SNMP gene family. Insect Biochem Mol Biol. 39(7):448–456.

Vogt RG, Riddiford LM. 1981. Pheromone binding and inactivation by moth antennae. Nature 293(5828):161–163.

Wanner KW, Isman MB, Feng Q, Plettner E, Theilmann DA. 2005. Developmental expression patterns of four chemosensory protein genes from the Eastern spruce budworm, *Chroistoneura fumiferana*. Insect Mol Biol. 14(3):289–300.

Wei H-S, Li K-B, Zhang S, Cao Y-Z, Yin J. 2017. Identification of candidate chemosensory genes by transcriptome analysis in *Loxostege sticticalis* Linnaeus. PLoS One 12(4):e0174036.

Wu C, et al. 2016. De novo transcriptome analysis of the common New Zealand stick insect *Clitarchus hookeri* (Phasmatodea) reveals genes involved in olfaction, digestion and sexual reproduction. Hull, JJ, editor. PLoS One 11(6):e0157783.

Zhang YV, Ni J, Montell C. 2013. The molecular basis for attractive salt-taste coding in *Drosophila*. Science 340(6138):1334–1338.

Zhou J-J, et al. 2010. Genome annotation and comparative analyses of the odorant-binding proteins and chemosensory proteins in the pea aphid *Acyrthosiphon pisum*. Insect Mol Biol. 19:113–122.

Associate editor: Mar Alba

Comparative genomics reveals thousands of novel chemosensory genes and massive changes in chemoreceptor repertories across chelicerates

Vizueta J., Rozas J. and Sánchez-Gracia A.

Supplementary Material

Table S1A. Summary of CS coding sequences identified in this study

Species	TOTAL				GR				IR/iGluR				OR				CSP			
			Not annotated in				Not annotated in				Not annotated in				Not annotated in				Not annotated in	
	Total	Annot.	GFF	Genome	Total	Annot.	GFF	Genome	Total	Annot.	GFF	Genome	Total	Annot.	GFF	Genome	Total	Annot.	GFF	Genome
D. melanogaster	279	279	0	0	60	60	0	0	80	80	0	0	60	60	0	0	4	4	0	0
D. pulex	243	154	89	0	58	58	0	0	161	93	68	0	0	0	0	0	3	3	0	0
S. maritima	220	147	72	21	76	76	0	0	105	69	36	20	0	0	0	0	2	2	0	0
Li. polyphemus	238	0	105	133	68	0	2	66	137	0	80	57	0	0	0	0	0	0	0	0
T. urticae	352	709	59	8	265	689*	0	0	20	20	0	0	0	0	0	0	0	0	0	0
M. occidentalis	131	139	34	4	18	64*	0	0	75	75	0	0	0	0	0	0	0	0	0	0
I. scapularis	196	92	58	46	62	62	0	0	104	29	34	41	0	0	0	0	1	1	0	0
C. exilicauda	1681	0	137	1544	1234	0	24	1210	414	0	88	326	0	0	0	0	0	0	0	0
M. martensii	891	0	117	774	673	0	23	650	186	0	76	110	0	0	0	0	0	0	0	0
A. geniculata	232	0	49	183	137	0	2	135	59	0	25	34	0	0	0	0	0	0	0	0
S. mimosarum	338	0	78	260	229	0	4	225	72	0	43	29	0	0	0	0	0	0	0	0
La. hesperus	184	0	49	135	76	0	0	76	70	0	30	40	0	0	0	0	0	0	0	0
P. tepidariorum	1016	0	209	807	819	0	52	767	144	0	105	39	0	0	0	0	0	0	0	0
Lo. reclusa	297	0	60	237	203	0	1	202	62	0	44	18	0	0	0	0	0	0	0	0
Total chelicerata	6026	940	955	4131	4254	815	108	3331	1343	124	525	694	0	0	0	0	1	1	0	0

* Proteins already characterized in previous studies

** Annot., Proteins that were already annotated as CS in genome data

** Not annotated in GFF, genes with predicted proteins but without functional annotation in genome data (uncharacterized peptides)

** Not annotated in Genome, genes undetected in genome sequences (completely new genes identified here using TBLASTN searches)

Species	OBP				OBP-like				NPC2				CCP				CD36-SNMP			
			Not annotated in				Not annotated in				Not annotated in				Not annotated in				Not annotated in	
	Total	Annot.	GFF	Genome	Total	Annot.	GFF	Genome	Total	Annot.	GFF	Genome	Total	Annot.	GFF	Genome	Total	Annot.	GFF	Genome
D. melanogaster	52	52	0	0	0	0	0	0	9	9	0	0	0	0	0	0	14	14	0	0
D. pulex	0	0	0	0	0	0	0	0	13	0	13	0	0	0	0	0	8	0	8	0
S. maritima	0	0	0	0	21	0	21	0	8	0	7	1	0	0	0	0	8	0	8	0
Li. polyphemus	0	0	0	0	2	0	2	0	14	0	10	4	0	0	0	0	17	0	11	6
T. urticae	0	0	0	0	4	0	4	0	48	0	42	6	0	0	0	0	15	0	13	2
M. occidentalis	0	0	0	0	4	0	4	0	14	0	13	1	0	0	0	0	20	0	17	3
I. scapularis	0	0	0	0	3	0	3	0	16	0	15	1	0	0	0	0	10	0	6	4
C. exilicauda	0	0	0	0	2	0	2	0	15	0	11	4	1	0	1	0	15	0	11	4
M. martensii	0	0	0	0	2	0	2	0	15	0	9	6	0	0	0	0	15	0	7	8
A. geniculata	0	0	0	0	1	0	1	0	20	0	14	6	0	0	0	0	15	0	7	8
S. mimosarum	0	0	0	0	4	0	4	0	16	0	13	3	7	0	7	0	10	0	7	3
La. hesperus	0	0	0	0	2	0	2	0	15	0	2	13	7	0	7	0	14	0	8	6
P. tepidariorum	0	0	0	0	4	0	4	0	20	0	19	1	21	0	21	0	8	0	8	0
Lo. reclusa	0	0	0	0	2	0	2	0	18	0	7	11	2	0	2	0	10	0	4	6
Total chelicerata	0	0	0	0	30	0	30	0	211	0	155	56	38	0	38	0	149	0	99	50

Table S1B. Number of complete and incomplete copies and pseudogenes

Species	Total			GR			IR/iGluR			CSP			OBP-like			NPC2			CCP			CD36-SNMP		
	CP	IP	Ψ	CP	IP	Ψ	CP	IP	Ψ	CP	IP	Ψ	CP	IP	Ψ	CP	IP	Ψ	CP	IP	Ψ	CP	IP	Ψ
D. melanogaster	167	0	0	60	0	0	80	0	0	4	0	0	0	0	0	9	0	0	0	0	0	14	0	0
D. pulex	165	69	9	56	0	2	86	68	7	3	0	0	0	0	0	13	0	0	0	0	0	7	1	0
S. maritima	164	23	33	61	0	15	66	21	18	2	0	0	21	0	0	7	1	0	0	0	0	7	1	0
L. polyphemus	103	104	31	37	23	8	47	71	19	0	0	0	2	0	0	11	0	3	0	0	0	6	10	1
T. urticae	525	29	222	447	22	220	18	2	0	0	0	0	4	0	0	44	3	1	0	0	0	12	2	1
M. occidentalis	161	6	10	58	0	6	73	0	2	0	0	0	4	0	0	13	0	1	0	0	0	13	6	1
I. scapularis	107	77	12	57	0	5	29	68	7	1	0	0	3	0	0	15	1	0	0	0	0	2	8	0
C. exilicauda	557	501	623	518	338	378	25	148	241	0	0	0	2	0	0	7	5	3	1	0	0	4	10	1
M. martensii	202	359	330	173	242	258	19	99	68	0	0	0	2	0	0	7	4	4	0	0	0	1	14	0
A. geniculata	82	128	22	54	65	18	7	49	3	0	0	0	1	0	0	13	6	1	0	0	0	7	8	0
S. mimosarum	166	134	38	120	81	28	13	50	9	0	0	0	4	0	0	14	1	1	6	1	0	9	1	0
L. hesperus	67	101	16	37	32	7	10	53	7	0	0	0	2	0	0	8	5	2	7	0	0	3	11	0
P. tepidariorum	513	348	155	383	288	148	81	56	7	0	0	0	4	0	0	17	3	0	20	1	0	8	0	0
L. reclusa	163	108	26	148	34	21	1	60	1	0	0	0	2	0	0	9	6	3	2	0	0	1	8	1
Total chelicerata	2646	1895	1485	2032	1125	1097	323	656	364	1	0	0	30	0	0	158	34	19	36	2	0	66	78	5

Table S1C. Summary of the CS family sizes

Species	N50	Scaffolds	Predicted Proteins	Total			GR			IR/iGluR								
										IR + iGluR			IR			iGluR		
				S_{CP}	S_{MIN}	S_{MAX}	S_{CP}	S_{MIN}	S_{MAX}	S_{CP}	S_{MIN}	S_{MAX}	S_{CP}	S_{MIN}	S_{MAX}	S_{CP}	S_{MIN}	S_{MAX}
D. melanogaster	-	-	-	167	167	167	60	60	60	80	80	80	66	66	66	14	14	14
D. pulex	642,089	5,186	31,524	165	233	234	56	56	56	86	153	154	79	146	147	7	7	7
S. maritima	139,451	14,739	15,012	164	187	187	61	61	61	66	87	87	54	75	75	12	12	12
Li. polyphemus	254,089	286,792	23,300	103	185	207	37	58	60	47	104	118	19	75	87	28	29	31
T. urticae	2,993,488	640	18,342	525	552	554	447	469	469	18	19	20	5	6	7	13	13	13
M. occidentalis	896,831	2,210	11,738	161	165	167	58	58	58	73	73	73	58	58	58	15	15	15
I. scapularis	76,228	369,492	21,186	107	164	184	57	57	57	29	82	97	15	68	85	14	14	12
C. exilicauda	342,549	10,457	30,465	557	1003	1058	518	832	856	25	150	173	15	135	158	10	15	15
M. martensii	45,228	92,408	32,016	202	520	561	173	399	415	19	101	118	10	87	103	9	14	15
A. geniculata	480,636	68,653	26,888	82	182	210	54	115	119	7	38	56	1	26	44	6	12	12
S. mimosarum	13,889	151,814	17,364	166	283	300	120	200	201	13	47	63	6	34	49	7	13	14
La. hesperus	465,572	31,445	32,186	67	138	168	37	67	69	10	43	63	9	38	58	1	5	5
P. tepidariorum	63,233	143,678	20,617	513	824	861	383	643	671	81	128	137	62	108	117	19	20	20
Lo. reclusa	20,294	4,986,575	76,238	163	239	271	148	176	182	1	40	61	0	29	50	1	11	11
Total chelicerata				2646	4255	4541	2032	3074	3157	323	825	979	200*	664	816	123	161	163

** The 249 complete IR described in the text (named t-IR) are these 200 copies plus the 49 with the LBD domain: the ligand channel domain (Lig_chan; PF00060).*

S_{MIN}: Complete proteins + minimum number of partial proteins that can be unequivocally assigned to different proteins (IPmin)

S_{MAX}: Complete proteins + total partial proteins (IP)

Species	CSP			OBP-like			NPC2			CCP			CD36-SNMP		
	S_{CP}	S_{MIN}	S_{MAX}	S_{CP}	S_{MIN}	S_{MAX}	S_{CP}	S_{MIN}	S_{MAX}	S_{CP}	S_{MIN}	S_{MAX}	S_{CP}	S_{MIN}	S_{MAX}
D. melanogaster	4	4	4	0	0	0	9	9	9	0	0	0	14	14	14
D. pulex	3	3	3	0	0	0	13	13	13	0	0	0	7	8	8
S. maritima	2	2	2	21	21	21	7	8	8	0	0	0	7	8	8
Li. polyphemus	0	0	0	2	2	2	11	11	11	0	0	0	6	10	16
T. urticae	0	0	0	4	4	4	44	47	47	0	0	0	12	13	14
M. occidentalis	0	0	0	4	4	4	13	13	13	0	0	0	13	17	19
I. scapularis	1	1	1	3	3	3	15	16	16	0	0	0	2	5	10
C. exilicauda	0	0	0	2	2	2	7	11	12	1	1	1	4	7	14
M. martensii	0	0	0	2	2	2	7	10	11	0	0	0	1	8	15
A. geniculata	0	0	0	1	1	1	13	19	19	0	0	0	7	9	15
S. mimosarum	0	0	0	4	4	4	14	15	15	6	7	7	9	10	10
La. hesperus	0	0	0	2	2	2	8	12	13	7	7	7	3	7	14
P. tepidariorum	0	0	0	4	4	4	17	20	20	20	21	21	8	8	8
Lo. reclusa	0	0	0	2	2	2	9	15	15	2	2	2	1	4	9
Total chelicerata	1	1	1	30	30	30	158	189	192	36	38	38	66	98	144

Table S2. Classification of IR/iGluR subfamilies

Subfamily	Subgroup	Members
NMDA	NMDAR1	LpolIR16, LpolIR29, LpolIR41, LpolIR8, TurtIR17, TurtIR18, CexiIR63p, DmelNMDAR1, Dpul1255143, IscaNMDAR01, LrecIR1, MmarIR13, PtepIR34, SmarNMDAR1, SmimIR5
NMDA	NMDAR2	LpolIR39, LpolIR5, LpolIR54, LpolIR7, PtepIR37, PtepIR76, PtepIR93, SmimIR3, SmimIR6, SmimIR7, AgenIR20p, LhesIR35p, MmarIR144p, AgenIR32p, CexiIR56p, DmelNMDAR2, IscaNMDAR02, LrecIR29p, MmarIR11, CexiIR51p, MmarIR143p, SmarIR101p, CexiIR52p, Dpul141686, LrecIR33p, SmarNMDAR2, TurtIR2
NMDA	NMDAR3	LpolIR33, LpolIR42, LpolIR44, LpolIR53, MmarIR19, MmarIR7, PtepIR60, PtepIR80, AgenIR15p, LhesIR2, SmimIR61p, AgenIR28p, CexiIR78p, IscaNMDAR03, LrecIR6p, SmarNMDAR3, TurtIR15, CexiIR3, SmimIR70p, LrecIR46p
KAINATE	KAINATE	AgenIR1, AgenIR2, AgenIR5, AgenIR6, AgenIR7, CexiIR1, CexiIR15, CexiIR11, CexiIR12, CexiIR26, DmelCG11155, DmelCG3822, DmelCG5621, DmelCG9935, Dmelclumsy, DmelGluRIID, DmelGluRIIE, DmelGluRIIA, DmelGluRIIC, DmelGluRIIB, Dpul1309661, Dpul1327119, Dpul155220, Dpul155388, IscaIR381p, IscaKA05, IscaKA06, IscaKA03, IscaKA04, LpolIR109p, LpolIR13, LpolIR23, LpolIR25, LpolIR27, LpolIR32, LpolIR48, LpolIR6, LpolIR12, LpolIR46, LpolIR51, LpolIR40, MmarIR1, MmarIR12, PtepIR137p, PtepIR30, PtepIR49, PtepIR52, PtepIR68, PtepIR69, PtepIR70, PtepIR73, PtepIR31, PtepIR39, PtepIR53, SmarKA01, SmarKA02, SmarKA03, SmarKA04, SmarKA05, SmimIR10, SmimIR12, SmimIR11, SmimIR4, TurtIR10, TurtIR11, TurtIR13, TurtIR7, TurtIR8, TurtIR9, LhesIR4, MocclR103, LrecIR2, AgenIR9, SmimIR9, MmarIR173p, MmarIR174p, CexiIR2, MmarIR15, MocclR101, LrecIR10p, LrecIR15p, LrecIR24p, SmimIR13
AMPA	AMPA	CexiIR17, CexiIR25, DmelGluRIA, DmelGluRIB, IscaAMPAR01, IscaAMPAR02, LpolIR21, LpolIR22, LpolIR26, LpolIR35, LpolIR52, PtepIR51, PtepIR67, PtepIR81, SmarAMPA02, SmarAMPA03, TurtIR12, TurtIR3, TurtIR5, AgenIR21p, LhesIR38p, LrecIR8p, SmarAMPA01, SmimIR2, AgenIR24p, LhesIR33p, LrecIR20p, Dpul1326779, LrecIR19p, MmarIR6, CexiIR4, MocclR104, MmarIR142p
IR25a/IR8a	IR25a	LpolIR47, LpolIR49, AgenIR42p, LrecIR51p, SmimIR34p, CexiIR10, DmelIR25a, DpulIR25a, IscaIR25a, LhesIR24p, MmarIR148p, MocclR25a, PtepIR97, SmarIR25a, TurtIR16, TurtIR14
IR25a/IR8a	IR8a	DmelIR8a, LpolIR11, SmarIR8a
IR	IR93a	CexiIR18, DmelIR93a, DpulIR93a, IscaIR93a, LhesIR31p, LpolIR50, LrecIR30p, MmarIR16, MocclR108, PtepIR96, TurtIR1

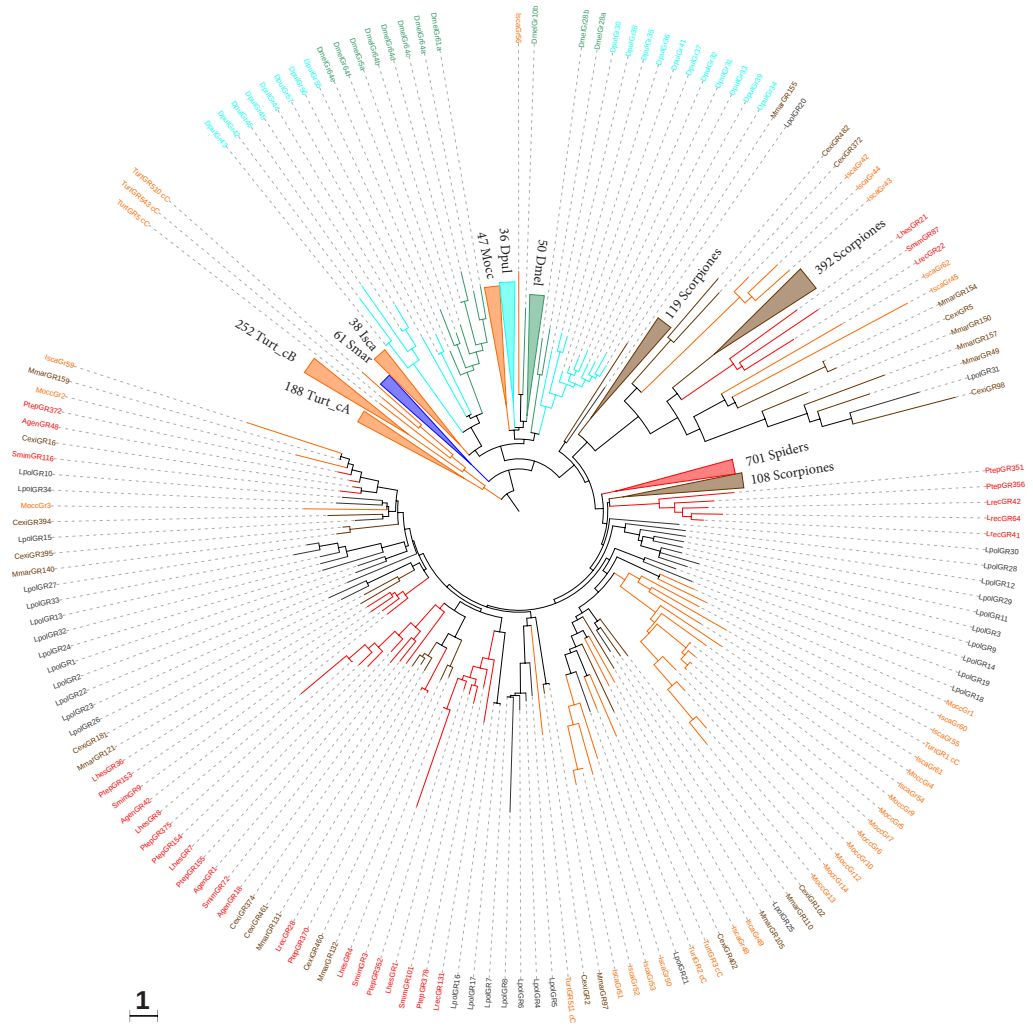


Figure S1. Phylogenetic tree of the GR family members across arthropods. The different species are depicted in colors as in Fig. 1. Monophyletic clades with 30 or more sequences from the same group are collapsed. Actual number of collapsed branches per clade is indicated in each case. The scale bar represents 1 amino acid substitution per site.

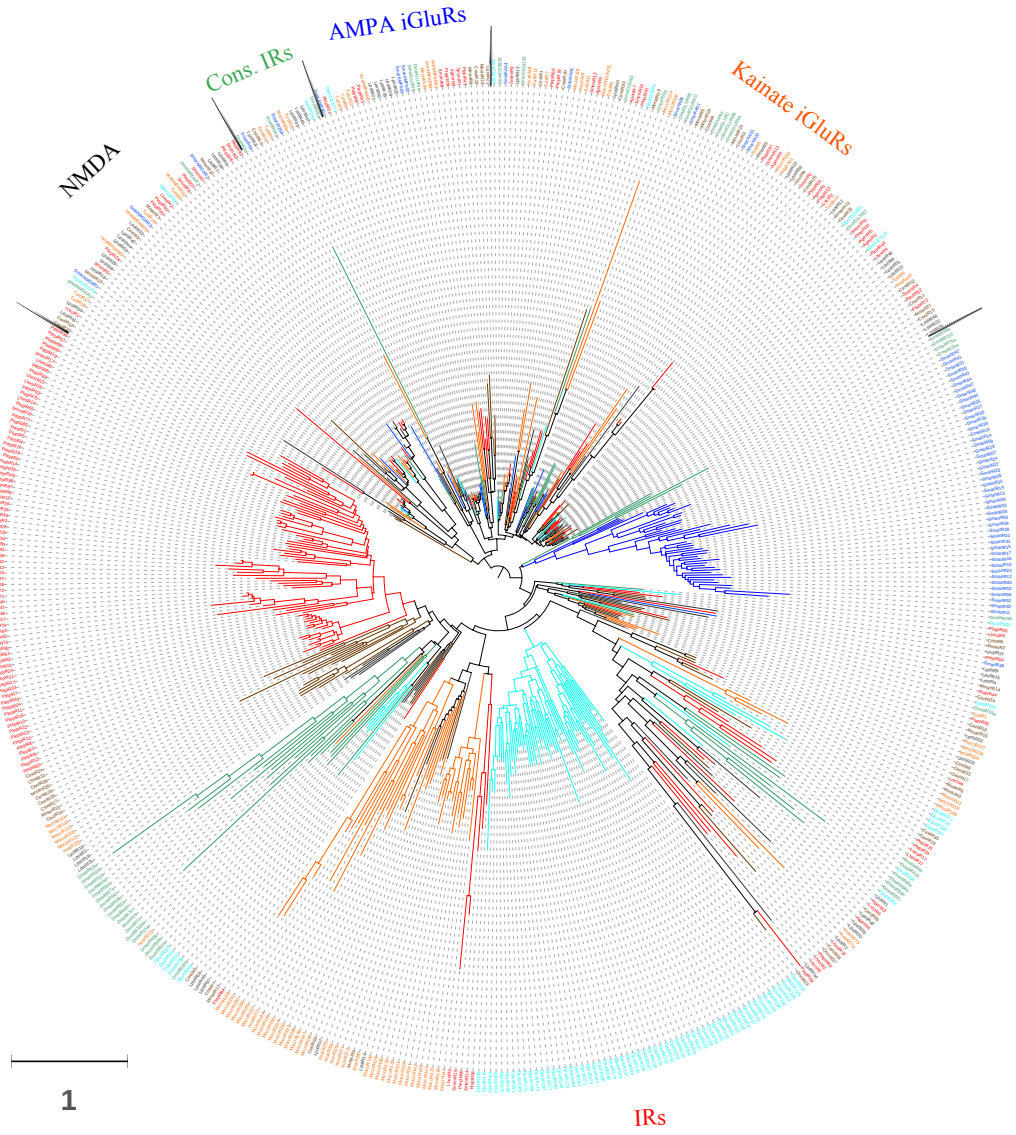


Figure S2. Phylogenetic tree of the IR/iGluR family members across arthropods as presented in Fig. 3 but including all sequence names.

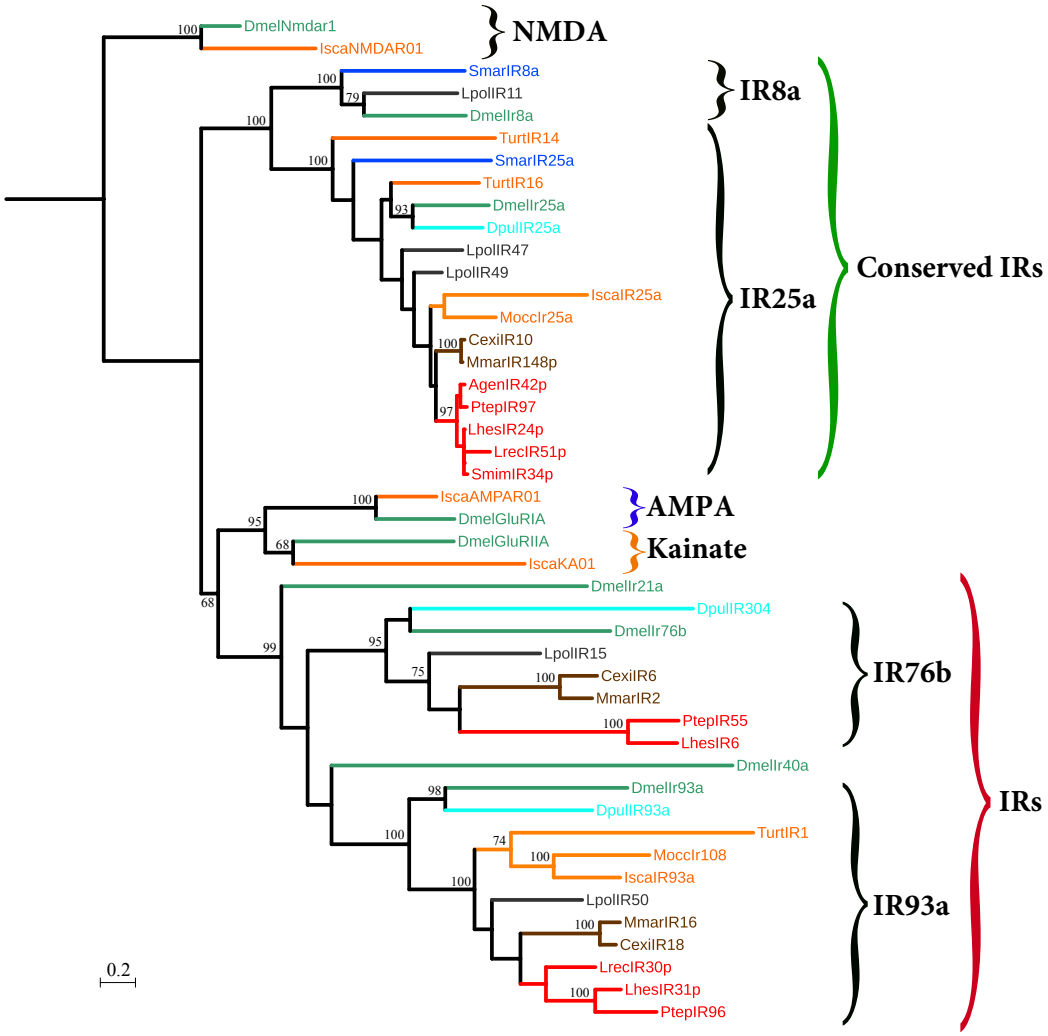


Figure S3. Phylogenetic tree of the members of the Ir subfamily conserved across arthropods. Two proteins representing the three other main groups of iGluRs (one from *I. scapularis* and the other from *D. melanogaster)* are also included in the tree. The phylogenetic analysis was performed with RAxM with 500 bootstrap replicates, indicated in the corresponding branches. Lineages and species names are coloured as in Fig. 1. The scale bar represents 0.2 amino acid substitutions per site.

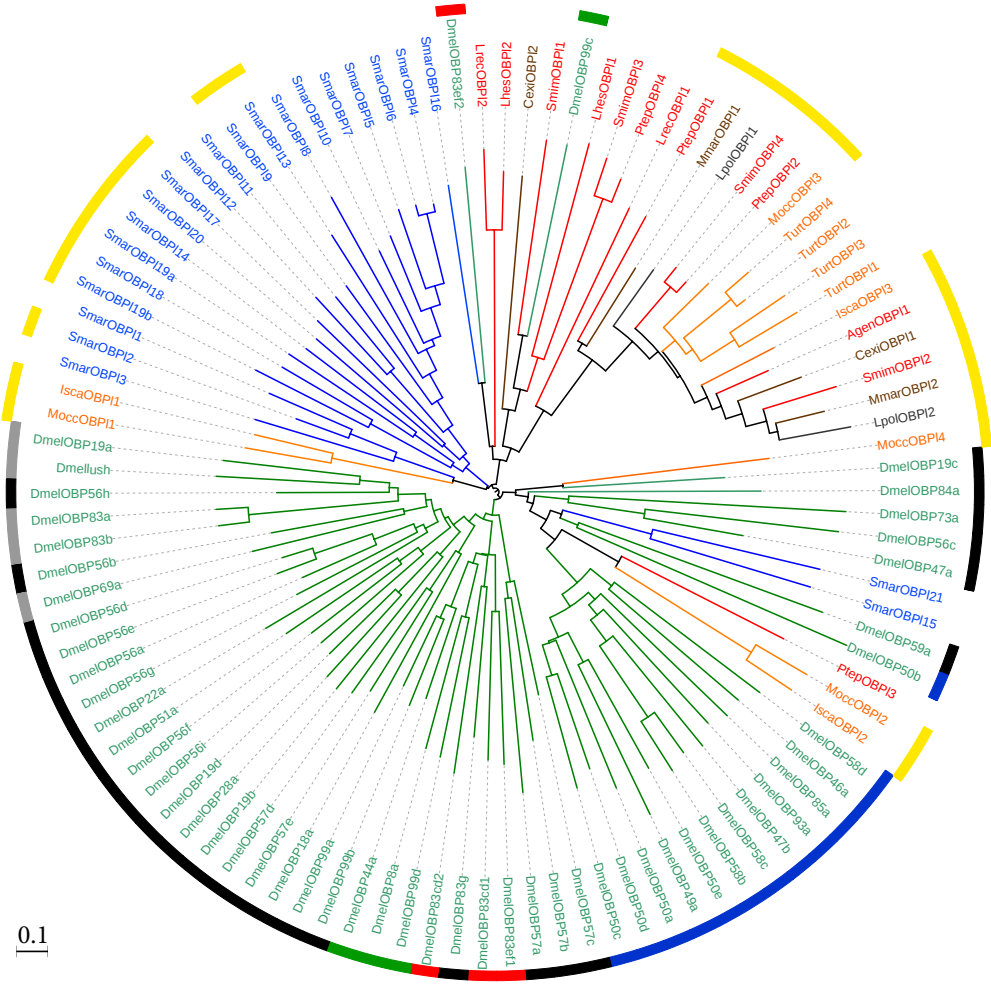


Figure S4. Phylogenetic relationships of the OBP-*like* and some insect (*D. melanogaster*) OBP family members. Lineages and species names are coloured as in Figure 1. Each colour of the inner circle identifies an OBP subfamily. Black, Classic; green, Minus-C; blue, Plus-C; red, Dimer; gray, ABPII. The outer yellow circle identifies the OBP-*like* sequences from non-insect species with a PBP/GOBP domain (IPR006170). The scale bar represents 0.1 amino acid substitutions per site.

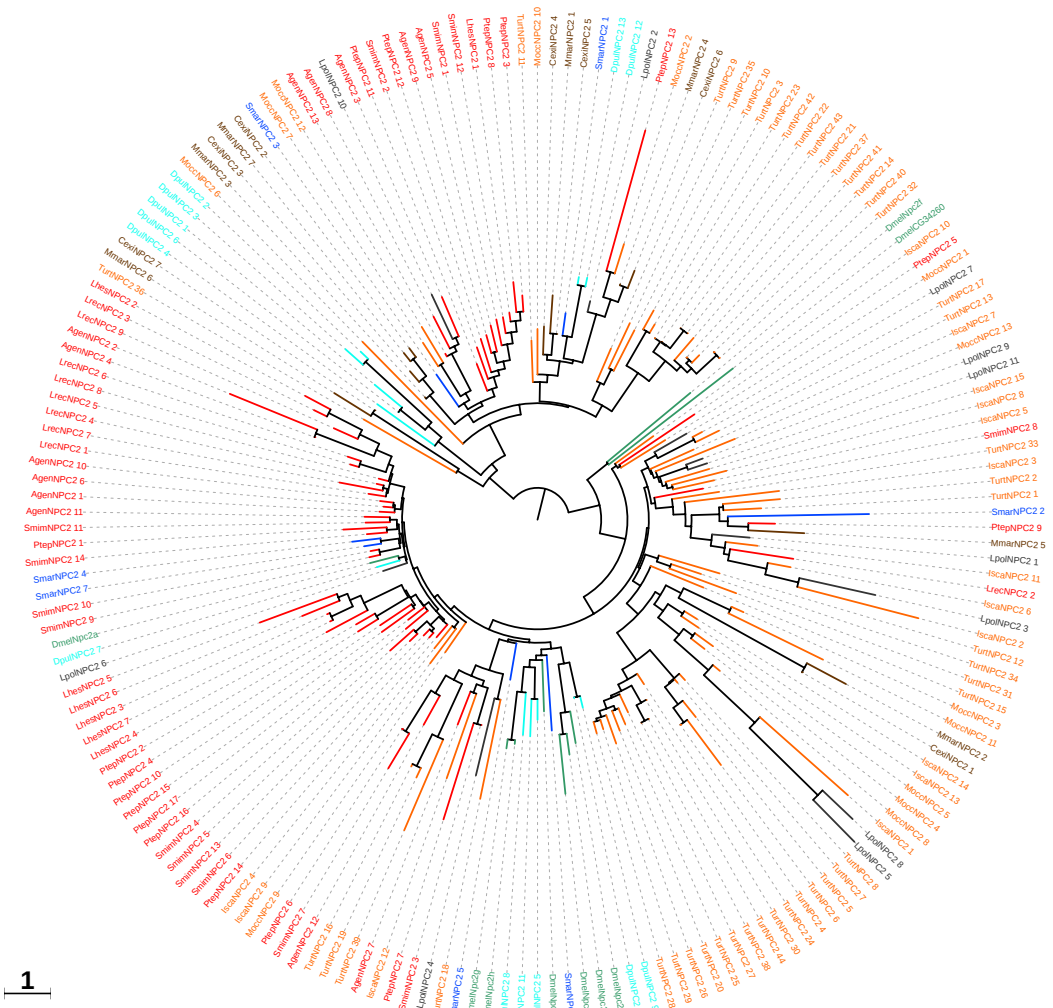


Figure S5. Phylogenetic tree of NPC2 proteins. Lineages and species names are coloured as in Fig. 1. The scale bar represents 1 amino acid substitution per site

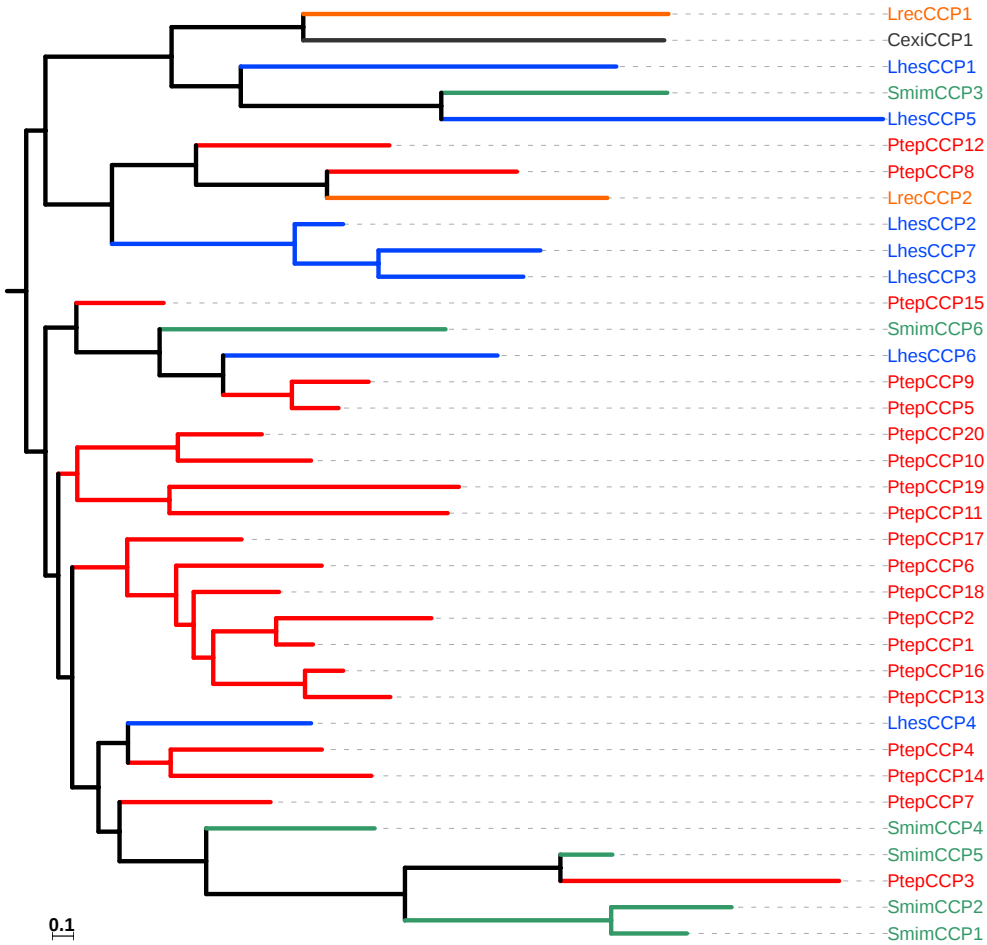


Figure S6. Phylogenetic tree of CCP family members in spiders and scorpions. The spider lineages and sequence names are shadowed in different colors. Brown, *C. exilicauda*; green, *S. mimosarum*; blue, *La. Hesperus*; red, *P. tepidariorum*; orange, *Lo. reclusa*. The scale bar represents 0.1 amino acid substitutions per site.

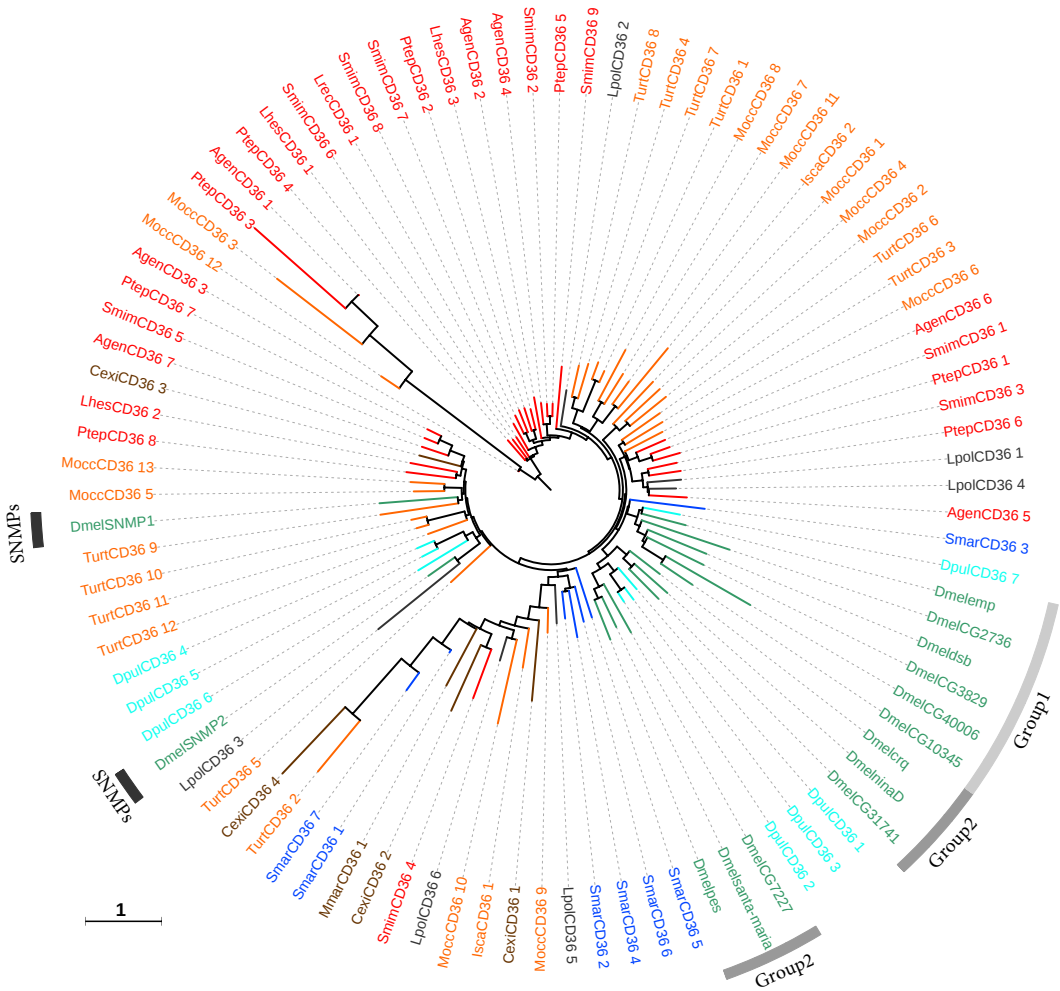


Figure S7. Phylogenetic tree of the CD36/SNMP family members across arthropods. Lineages and species names are colored as in Fig. 1. The outer circle identifies the phylogenetic subgroups described in *D. melanogaster*. The scale bar represents 1 amino acid substitution per site.

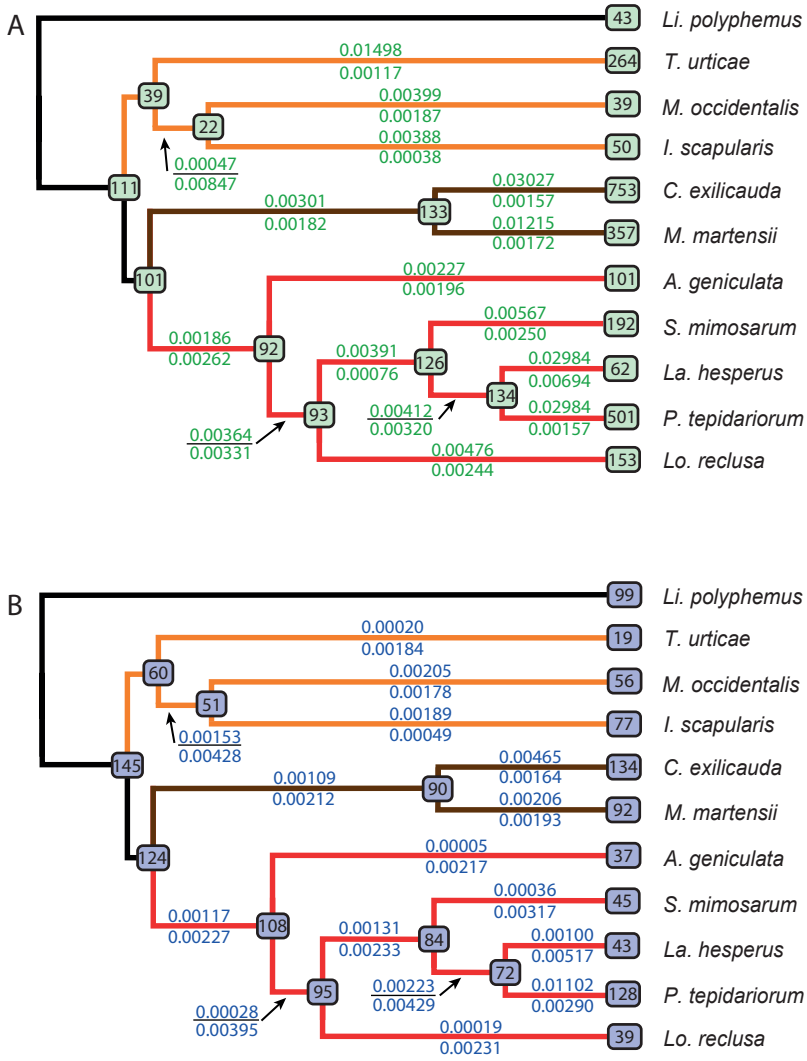


Figure S8. Gene turnover rates of chemoreceptors across chelicerates. (*A*) GR family. (*B*) IR/iGluR family. Estimates obtained from the dataset used to estimate S_{MIN}. Numbers above and below each branch indicate lineage-specific gene birth and death rates (events per gene per million year), respectively. Estimates in very short and outgroup branches have large uncertainty and are not showed. Numbers in the ancestral nodes show the estimated family sizes. Numbers at the tips indicate the number of sequences used for the analysis; such values can differ from S_{MIN} since only sequences that clustered in an orthogroup (with 3 or more sequences) were included in the analysis.

4

Chance and predictability in evolution: the genomic basis of convergent dietary specializations in an adaptive radiation

La coexistencia de múltiples fenotipos en organismos con un origen independiente hace que las radiaciones adaptativas en islas sean laboratorios naturales para estudiar la convergencia y el paralelismo en la evolución. En la radiación de las arañas del género *Dysdera* en las Islas Canarias, la diversificación de especies ha tenido lugar junto con eventos repetidos de especialización trófica. Estos cambios en la dieta afectan especialmente a la alimentación específica de isópodos, y están acompañados por modificaciones morfológicas (principalmente en los quelíceros), fisiológicas y en el comportamiento. Para comprender las bases moleculares de esta radiación adaptativa, hemos realizado un análisis exhaustivo de transcriptómica comparativa en cinco especies de *Dysdera* endémicas de las Islas Canarias, que representan dos eventos evolutivos geográficamente independientes de especialización trófica. Tras controlar por los posibles efectos de hemiplasia, nuestros análisis de expresión diferencial y constricción selectiva identificaron varios cambios genéticos que podrían estar asociados en la adaptación a la alimentación especializada de isópodos, incluyendo genes relacionados con la detoxificación y homeostasis de metales pesados, el metabolismo de nutrientes y venenos. Por lo tanto, nuestros resultados han proporcionado nuevos conocimientos sobre la base genómica de un evento de convergencia en el cambio de dieta asociado a la diversificación de estas especies. A su vez, hemos identificado cambios genómicos a distintos niveles jerárquicos, potencialmente producidos por cambios evolutivos convergentes, incluyendo genes específicos, genes con funciones similares e incluso posiciones aminoacídicas puntuales. En términos generales, este estudio profundiza en el conocimiento de las radiaciones adaptativas y la predictibilidad en la evolución.

Received: 21 December 2018 | Revised: 18 July 2019 | Accepted: 19 July 2019

DOI: 10.1111/mec.15199

ORIGINAL ARTICLE

MOLECULAR ECOLOGY WILEY

Chance and predictability in evolution: The genomic basis of convergent dietary specializations in an adaptive radiation

Joel Vizueta[1] | Nuria Macías-Hernández[2,3] | Miquel A. Arnedo[4] | Julio Rozas[1] | Alejandro Sánchez-Gracia[1]

[1]Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain

[2]Laboratory for Integrative Biodiversity Research, Finnish Museum of Natural History, University of Helsinki, Helsinki, Finland

[3]Island Ecology and Evolution Research Group, Instituto de Productos Naturales y Agrobiología (IPNA-CSIC), Tenerife, Spain

[4]Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals, Facultat de Biologia, Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain

Correspondence
Julio Rozas and Alejandro Sánchez-Gracia, Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Diagonal 643, 08028, Barcelona, Spain.
Emails: jrozas@ub.edu; elsanchez@ub.edu

Funding information
Comissió Interdepartamental de Recerca i Innovació Tecnològica, Grant/Award Number: 2014SGR1055, 2014SGR1604, 2017SGR1287 and 2017SGR83; Secretaría de Estado de Investigación, Desarrollo e Innovación, Grant/Award Number: CGL2012-36863, CGL2013-45211, CGL2016-75255 and CGL2016-80651; Ministerio de Economía y Competitividad, Grant/Award Number: BES-2014-068437

Abstract

The coexistence of multiple eco-phenotypes in independently assembled communities makes island adaptive radiations the ideal framework to test convergence and parallelism in evolution. In the radiation of the spider genus *Dysdera* in the Canary Islands, species diversification occurs concomitant with repeated events of trophic specialization. These dietary shifts, to feed primarily on woodlice, are accompanied by modifications in morphology (mostly in the mouthparts), behaviour and nutritional physiology. To gain insight into the molecular basis of this adaptive radiation, we performed a comprehensive comparative transcriptome analysis of five Canary Island *Dysdera* endemics representing two evolutionary and geographically independent events of dietary specialization. After controlling for the potential confounding effects of hemiplasy, our differential gene expression and selective constraint analyses identified a number of genetic changes that could be associated with the repeated adaptations to specialized diet of woodlice, including some related to heavy metal detoxification and homeostasis, the metabolism of some important nutrients and venom toxins. Our results shed light on the genomic basis of an extraordinary case of dietary shift convergence associated with species diversification. We uncovered putative molecular substrates of convergent evolutionary changes at different hierarchical levels, including specific genes, genes with equivalent functions and even particular amino acid positions. This study improves our knowledge of rapid adaptive radiations and provides new insights into the predictability of evolution.

KEYWORDS
comparative transcriptomics, diet specialization, differential gene expression, heavy metals, oceanic islands, phenotypic convergence, positive selection, spiders, toxins

1 | INTRODUCTION

The current limited knowledge of the evolutionary mechanisms underlying diversification compromises our ability to manage and conserve biodiversity (Mergeay & Santamaria, 2012). Evolutionary biology provides a unifying conceptual framework to successfully identify key diversification drivers through the study of molecular variation. As many other fields, evolutionary biology has fully entered the genomics era, which opens up the possibility of tackling long-standing questions regarding biodiversity in a more fruitful way and at a lower cost (Losos et al., 2013). Although often seen as a gradual process that requires the action of different evolutionary forces

acting steadily over long periods of time (Coyne & Orr, 2004), speciation can be very rapid under unstable environmental and ecological conditions. In fact, one of the most promising approaches to disclose the relative impact of these driving forces is the study of species radiations in nature, that is the rapid appearance of a high number of species from a single common ancestor (Schluter, 2000). In adaptive radiations, such as the classic examples of Darwin's finches (Almén et al., 2016) and the cichlids in the great lakes of Eastern Africa (Henning & Meyer, 2014), significant morphological differences appear over short periods of time despite the low levels of genetic divergence accumulated at the genomic level. Nevertheless, the relative role of natural selection and of other nonadaptive forces in such relevant evolutionary processes is a matter of scientific debate (Muschick, Indermaur, & Salzburger, 2012).

Oceanic islands are considered natural laboratories for studying evolution. The entire biota of these islands is derived from a few initial colonization events followed by local diversification, which generates high levels of endemism and ecomorphological differentiation (MacArthur & Wilson, 1967; Mayr, 1942; Whittaker & Fernández-Palacios, 2007). Thus, the biota of oceanic islands can be interpreted as the result of successful independent evolutionary experiments starting with a single or multiple colonization events from the continent (Emerson, 2002). The comparative analysis of such independent events and the subsequent island radiation (both within and between islands) in different archipelagos provides new insights into the general evolutionary process generating biological diversity (Gillespie & Roderick, 2002; Losos & Ricklefs, 2009). Such approximation has been successfully applied in a number of studies on oceanic islands (Losos, Jackman, Larson, Queiroz, & Rodriguez-Schettino, 1998; Stroud & Losos, 2016), such as Hawaii (Gillespie, 2004), the Galapagos (Grant & Grant, 2008) and the Canary Islands and Madeira archipelagos (Juan, Emerson, Oromí, & Hewitt, 2000; Machado, Rodríguez-Expósito, López, & Hernández, 2017), where explicit hypotheses on the evolutionary processes underlying radiations have been tested.

The radiation of the genus *Dysdera* Latreille, 1804 (Araneae: Dysderidae), in the Canary Islands is one of the most spectacular examples of island species diversification within spiders (Arnedo, 2001; Arnedo, Oromí, Múrria, Macías-Hernández, & Ribera, 2007). As many as 47 endemic species of this species-rich Mediterranean genus (approximately 250 species) have been reported in the Canary Islands (Macias-Hernandez, de la Cruz López, Roca-Cusachs, Oromí, & Arnedo, 2016; World Spider Catalog, 2019). The spiders of the genus *Dysdera* are active nocturnal hunters that spend the daytime in silk retreats and are usually found under stones, dead logs or leaf litter or even living in caves (Arnedo et al., 2007). This genus stands out among spiders in having evolved trophic specialization; that is, several species have been shown to feed preferably (facultatively or even obligatorily) on terrestrial woodlice (Crustacea: Isopoda; Řezáč & Pekár, 2007; Řezáč, Pekár, & Lubin, 2008), a prey rejected by most generalist predators (Pekár, Líznarová, & Řezáč, 2016). Available evidence suggests that prey specialization (i.e., stenophagy) has appeared several times, both on the continent and on the islands. Interestingly, the morphology of mouth parts predicts both dietary preferences and capture strategy (chelicerae used as pincers, forks or keys) and the frequency of captures among the specialists (Řezáč et al., 2008). All cheliceral types observed in continental species have also evolved repeatedly in the Canary Islands, suggesting that prey segregation is a major driving force of the spectacular diversification of the genus on the islands (Arnedo et al., 2007). Woodlice are a difficult prey for other arthropods because of their morphological, chemical and behavioural defences (Gorvett, 1956; Sutton, 1980). These defences comprise dorsally protective armour, gland secretions producing repulsive odours, indigestibility to many predators and behavioural patterns such as nocturnal activity, rolling into a ball or adhering to surfaces when threatened (Schmalfuss, 1984; Sutton, 1980). In addition, these organisms accumulate high concentrations of heavy metals from the soil, making them even more toxic to predators (Drobne, 1997). Consequently, woodlice are rarely eaten by generalist predators. Within arthropods, only spiders and ants have developed specialized strategies to feed on this prey (Dejean, 1997; Pekár et al., 2016). Nevertheless, despite all this morphological and experimental evidence, the genetic basis of this remarkable adaptation is completely unknown.

Moreover, the study of the molecular basis of such an extraordinary phenotypic convergence offers an opportunity to address the question of predictability and repeatability of the evolutionary process. Given that it is not possible to rerun the tape of evolution, the study of parallel evolutionary outcomes in different scenarios provides a fairly good framework to ascertain both to what extent similar molecular solutions have been exploited repeatedly, and which aspects are predictable at different hierarchical levels (i.e., at the nucleotide, gene, pathway or function level). Among *Dysdera* spiders, the specialized woodlice eaters (i.e., oniscophagous species) possess, in addition to the morphological modifications of chelicera, important behavioural and nutritional adaptations to feed on isopods (Hopkin & Martin, 1985; Řezáč & Pekár, 2007; Toft & Macías-Hernández, 2017). With the aim of understanding the genetic basis of these specific adaptations and to shed some light on the long-standing debate of how predictable is molecular evolution, we designed a case study that included adult individuals from two pairs of recently diverged endemic specialist–generalist species from the Canary Islands, likely representing two phylogenetically and geographically independent dietary shifts from a generalist ancestor. Our survey included the GV pair: *Dysdera gomerensis* Strand, 1911 (El Hierro), and. *D. verneaui* Simon, 1883 (Tenerife), the TB pair: *D. tilosensis* Wunderlich, 1992, and. *D. bandamae* Schmidt, 1973 (Gran Canaria), and a third generalist endemic species external to both pairs: *D. silvatica* (La Gomera; M. A. Arnedo pers. Comm.; Macías-Hernández, Oromí, & Arnedo, 2008; Vizueta et al., 2017), which was used as an outgroup (Figure 1). We compared the transcriptome profiles and the selective constraint patterns between specialists and generalists to identify the genomic regions responsible for the rapid dietary adaptation of *Dysdera* species in the Canary Islands. We studied transcriptomic data from adult individuals, and we were able to detect putative

adaptive changes associated with food detection and assimilation, including its digestive and metabolic aspects. True homoplasy can arise by evolving the same (or similar) trait from either a nonshared common ancestor (convergent evolution) or a shared ancestor but through evolutionarily independent events (parallel evolution). Here, we will refer to both cases with the general term of "convergence". We aimed to detect those evolutionary changes required to explain a repeated character state in the two specialist lineages, either a gene expression profile or a selective constraint pattern, matching phenotypic convergence. Nevertheless, both incomplete lineage sorting (ILS; Maddison, 1997) and species hybridization can produce fundamental discordances between gene trees and the species tree, a phenomenon commonly referred to as "hemiplasy" (Avise & Robinson, 2008), giving rise to the illusion of homoplasy and the erroneous inference of convergence (Mendes, Hahn, & Hahn, 2016; Wu, Kostyun, Hahn, & Moyle, 2018).

Here, and after controlling for the potential confounding effects of hemiplasy, we identified clear signals of homoplasy at different hierarchical levels likely attributable to adaptive convergence in specialist species. Noticeably, we even find signals of this adaptive process at the amino acid level. The repeated changes matching phenotypic convergence found in this study mostly affected genes and gene functions associated with the strategy of detoxifying heavy metals (and perhaps other toxic substances) accumulated by woodlice, to the enhanced assimilation of some nutrients and, to a lesser extent, to venom composition.

2 | MATERIAL AND METHODS

2.1 | Study design and sample materials

Our study design included two pairs of phylogenetically related *Dysdera* species endemic from the Canary Islands. Each pair of close relatives was composed of a generalist and a specialist (stenophagous) species regarding their diet and shared a generalist ancestor, which implies that at least two specialization events occurred independently during the divergence of these four species, one on each species pair (Figure 1). Both the phylogenomic analysis performed here and recent multilocus-based phylogenies including other endemic species of this genus (M. A. Arnedo, N. Macías-Hernández, & A. Enguídanos, unpublished results) indicate that *D. gomerensis* and *D. verneaui* are true sister taxa, while *D. tilosensis* and *D. bandamae* are very closely related, although it is difficult to know if they are each other closest relatives. Similarly, the ancestral state reconstruction supports that the ancestor of the complete *Canarian* radiation was a generalist, while *D. tilosensis* is a derived specialist from a generalist ancestor. For the case of *D. gomerensis*, this is much more difficult to establish because of the phylogenetic uncertainty, probably due to a very rapid radiation of these species group. In any case, this rapid radiation however makes that most candidate changes in the *D. gomerensis* lineage (see below) would be adaptations to stenophagy, independently of whether the ancestor was a complete generalist, or just a facultative intermediate.

The two specialists' species of our study show modifications in their mouthparts that have been associated with a preference for using isopods as a prey (Řezáč et al., 2008; Macías-Hernández et al., *in preparation*; see Figure 1). We collected 16 individuals of *Dysdera tilosensis* (10 males and 6 females) and 14 individuals of *D. bandamae* (5 males and 9 females) in Gran Canaria, and 12 males of *D. verneaui* in Tenerife and 15 females of *D. gomerensis* in El Hierro (Table S1). We also included in the analysis a fifth Canary Island endemic *Dysdera* species, the generalist *D. silvatica*, as an outgroup and to polarize the evolutionary changes in internal branches (Vizueta et al., 2017; Figure 1).

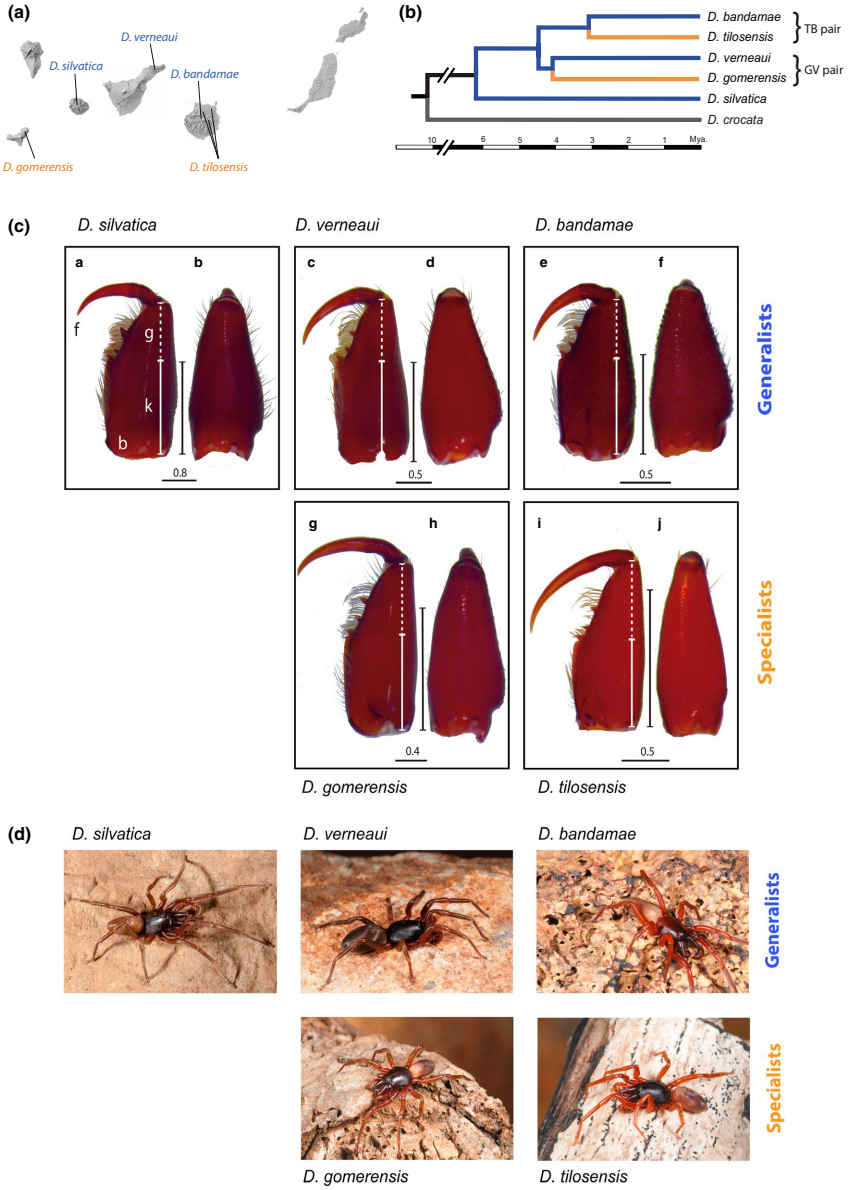
2.2 | Transcriptomic analysis

For each species, we sequenced the transcripts from the palps (*PALP*), the first pair of legs (*LEG#1*), all other legs (*LEG#234*), and the rest of the body (*REST*), separately in four different RNA-seq experiments. We applied this strategy to maximize the detection of low expressed genes, especially chemosensory gene family members in spider appendices (see Vizueta et al., 2017 and Frías-López et al., 2015; Appendices S1 and S2). Specimens were starved for two weeks at the laboratory and posteriorly fixed in liquid nitrogen and stored at −80°C until further processing. From the total RNA, we sequenced the transcriptomes in the Illumina HiSeq 4000 platform using pair-end libraries (100-bp reads; Table S1). A detailed description of raw data preprocessing, transcriptome assembly and functional annotation of the transcripts from the four species is available in Appendices S1 and S2.

2.3 | Species-tree, gene-tree discordance, and risk of hemiplasy

We identified all groups of homologous genes that share at least one member in the ancestor of the five *Dysdera* species (i.e.,

FIGURE 1 (a) Map of the Canary Islands showing the geographic location of capture localities. (b) Phylogenetic relationships and divergence times (scale bar) among surveyed *Dysdera* species. The continental species *D. crocata* was used to root the tree. (c) Dissecting scope images of the left chelicera: A–B: *Dysdera silvatica* female, La Gomera, A, ventral view; B, lateral view; C–D: *D. verneaui* female, Tenerife, C, ventral view, D, lateral view; E–F: *D. bandamae* female, Gran Canaria, E, ventral view, F, lateral view; G–H: *D. gomerensis* female, La Gomera, G, ventral view, H, lateral view; I–J: *D. tilosensis* male, Gran Canaria, I, lateral view, J, lateral view. Bars indicate the relative lengths of the different parts of the chelicerae to highlight differences between the standard (generalists) and elongated or slightly elongated (specialists) chelicerae. White bar: total length of the basal segment (b), dotted part: length of the cheliceral groove (g). Black bar: length of the cheliceral fang (f). In standard chelicerae, g is approximately 1/3 of b, and f is similar to the distance between the base of the segment and the end of the internal keel (k), while in elongated chelicerae, g is longer than 2/5 of f, and f is longer than k. Scale bar in mm. (d) Live images of the target *Dysdera* species; photo credit: Pedro Oromí [Colour figure can be viewed at wileyonlinelibrary.com]



orthology groups) using OrthoMCL with default parameters (Li, Stoeckert, & Roos, 2003). We further separated single-copy orthologs from multigene families. Since at the moment of starting this work, all published phylogenetic analyses including the studied species were based on few genes (Arnedo, 2001; Arnedo et al., 2007), and we performed a more comprehensive phylogenomic analysis using all single-copy orthologs across the five Canarian *Dysdera* species plus *D. crocata* Koch, 1839 (the phylogenetically closest continental species of this genus with available transcriptome data; Fernández, Hormiga, & Giribet, 2014; Figure 2). Only

complete or nearly complete transcripts free of premature stop codons were included in the analysis. The multiple sequence alignments (MSA) of the CDS of each orthology group were generated with the program T-Coffee (Notredame, Higgins, & Heringa, 2000) and further concatenated in a single MSA using in house Perl scripts. We set the GTRGAMMA substitution model in a partitioned scheme to obtain the maximum-likelihood (ML) tree in the software RAXML (Stamatakis, 2014). Model parameters were estimated independently for each single-copy ortholog, and node support was obtained after 500 bootstrap replicates.

We approximated the divergence times between the five Canarian *Dysdera* species by fitting the data from single-copy orthologs to the unrooted tree topology of the ML tree after excluding *D. crocata*. We set the same substitution model and partition scheme than in the previous RAXML analysis. We used the penalized likelihood method of Sanderson (2002), implemented in the program R8S v1.80, to generate the ultrametric tree and to estimate node ages

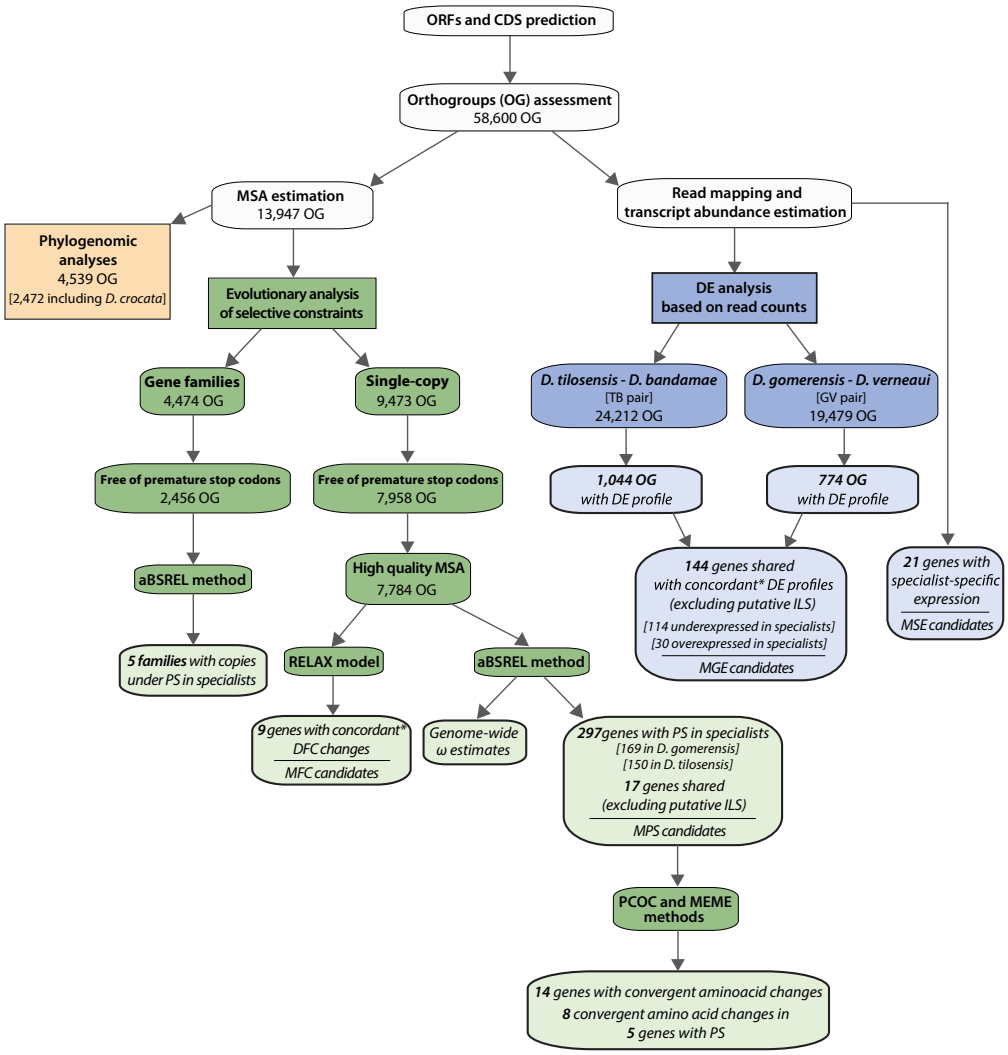


FIGURE 2 Core analyses workflow applied in this study, including a summary of the most relevant results. DE, differential expression; DFC, differential functional constraints; PS, positive selection; *, patterns matching the observed phenotypic convergence [Colour figure can be viewed at wileyonlinelibrary.com]

(Sanderson, 2003). We set a calibration point in the node representing the split of the *D. silvatica* lineage from the rest of lineages (3.4–7.8 Mya range; Macías-Hernández, Bidegaray-Batista, Emerson, Oromí, & Arnedo, 2013).

We also inferred a species tree that incorporates gene-tree uncertainty using ASTRAL (Zhang, Rabiee, Sayyari, & Mirarab, 2018). For that, we first estimated the ML tree of each individual MSA (i.e., a gene tree for each single-copy ortholog) with RAXML (setting the GTRGAMMA substitution model and calculating node support with 1,000 bootstrap replicates). Moreover, we estimated the hemiplasy risk factor (HRF) along the phylogeny using the PePo package (Guerrero & Hahn, 2018). For the analysis, we used the species tree inferred with ASTRAL (with branch lengths in $2N_e$ generation units), a very approximate estimate of the population scaled mutation rate in *D. silvatica* (θ = 0.011; estimate obtained from a short read alignment to the first genome draft of this species; Sánchez-Herrero et al., 2019), a generation time of 1.5 years, and six different effective population sizes, N_e (10^3, 5×10^3, 10^4, 5×10^4, 10^5 and 10^6). Finally, all candidate genes exhibiting resolved discordant topologies (i.e., with bootstrap support ≥75% in at least one node producing discordance with the species tree) were excluded for the downstream functional prediction analyses and their interpretation. Finally, we used the D_{FOIL} statistic (Pease & Hahn, 2015) to test for introgression between the specialist lineages in presence of ILS, using both *D. silvatica* and *D. crocata* as outgroups.

2.4 | Differential expression analyses

Differential expression (DE) analyses were performed separately in each generalist–specialist pair (GV and TB pairs; see Figure 1; Appendices S1 and S2). Raw reads of the RNA-seq from each species and body part were mapped back to their own reference CDS and to the CDS of the other species in the pair by using BOWTIE2 version 2.2.3 (Langmead & Salzberg, 2012). Read counts and TMM-normalized FPKMs (i.e., trimmed mean of log-expression ratios-normalized fragments per kb of exon per million reads mapped) were estimated for single-copy genes and multigene families using RSEM 1.2.19 software (Li & Dewey, 2011). To test for genes showing DE between specialists and generalist species, we calculated the negative binomial dispersion of read counts across species pairs of a set of housekeeping (HK) genes with EDGER version 3.18.1 (Robinson, McCarthy, & Smyth, 2010). We used this dispersion to conduct the DE analysis between specialist and generalist species. We merged all body parts (within a species) to homogenize the differences in the number of REST samples between species pairs. To avoid type I and II errors associated to this merging, especially when gene expression is higher in *REST* relative to legs (both *LEG#1* and *LEG#234*) and *PALP*, we used total read counts from all samples normalized for each library size to perform differential expression analyses. The *p*-values of these analyses (one per gene) were corrected for the false discovery rate (Benjamini & Hochberg, 1995; FDR). We considered that a gene is differentially expressed between two species when expression levels are significantly different with a FDR < 0.05.

2.5 | Selective constraints analyses

We used the adaptive Branch-Site Random Effects Likelihood (aBSREL) model implemented in the HyPhy package (Pond, Frost, & Muse, 2005; Smith et al., 2015) to test if positive selection has occurred repeatedly in the same gene in specialist lineages. This method is based on the parameter ω (the ratio of nonsynonymous (d_N) to synonymous (d_S) substitution rates, $\omega = d_N/d_S$) and allows fitting an optimal number of ω classes to codon sequence alignments of single-copy orthologs in each branch of the phylogeny (Figure 2; Appendices S1 and S2). Positive selection is inferred when a gene shows codons fitting a class with $\omega > 1$ in a particular lineage. We also tested for relaxation or intensification of the strength of natural selection in these single-copy orthologs in specialist lineages using the RELAX framework in HYPHY (Wertheim, Murrell, Smith, Kosakovsky Pond, & Scheffler, 2015). Besides, we applied the Mixed Effects Model of Evolution (MEME) implemented in the HYPHY package (Murrell et al., 2012) to identify individual sites evolving under episodic positive selection (in one or more lineages) in the set of candidates from PCOC analysis (see below). Both methods are based on the same principle of aBSREL of fitting different probabilistic models of the ω parameter distribution and also inferred positive selection when $\omega > 1$. Finally, we applied the aBSREL model to test for episodic positive selection acting on gene families in specialist lineages. In this case, we used the same workflow as for the single-copy orthologs but applying the FASTTREE program (Price, Dehal, & Arkin, 2010) to approximate a ML tree of each family.

2.6 | Convergent amino acid evolution

To detect convergent amino acid evolution in specialist lineages, we aligned the amino acid sequences of the PS candidates using the software PRANK and applied the method PCOC (Rey, Guéguen, Sémon, & Boussau, 2018; Profile Change with One Change), a recently developed approach to identify convergent shifts in the amino acid substitution rate across a phylogeny, to each individual MSA. Moreover, we used computer simulations to test the performance of PCOC method with our empirical data. We applied the same species tree, average sequence length and model parameters set in the PCOC analysis of the observed data to simulate sequences both with convergent (2% of sites undergoing convergent amino acid substitutions) and without convergent changes (Rey et al., 2018). Using these simulated sequences, we estimated the false discovery rate (FDR; using simulations without convergence) and true positive rate (TPR; using simulations with convergent amino acid substitutions) associated with this analysis.

2.7 | GO enrichment

We used R and GOSTATS (Falcon & Gentleman, 2007) to carry out the gene ontology (GO) enrichment analysis and REVIGO (Supek, Bošnjak, Škunca, & Šmuc, 2011) to generate a graphical representation of the results. We also used BLAST2GO suite (Conesa et al.,

2005) to identify KEGG pathways enriched in the list of candidates (Kanehisa & Goto, 2000). Hypergeometric tests were performed with dhyper function of the R package STATS.

3 | RESULTS

We constructed 16 RNA-seq data sets (four different body parts in four species) to obtain four new complete *Dysdera* transcriptomes (Table S1). As expected, both the number of species-specific transcripts (from 170,846 to 347,878) and the number of functionally annotated genes differed between species (Table 1), but the transcriptome completeness, measured as the number and integrity of CEG genes, was quite similar (Table S2). Only 30% of the transcripts encoded protein-coding genes; the rest corresponded to either noncoding transcripts or assembly artefacts (Table 1). Furthermore, ~35% of the predicted proteins showed no significant sequence similarity or conserved profiles with known arthropod genes (i.e., putative orphan genes of the *Dysdera* lineage). Among the annotated proteins, most were chelicerate specific, and ~66% of the top BLAST hits matched spider sequences (Figure S1).

We identified a total of 13,947 orthologous groups across the five Canarian *Dysdera* species, of which 7,958 were free of premature stop codons, and 4,539 showed complete sequences in all species (Figure 2). The number of single-copy orthologus across the five species was 9,473, a number that increased to 19,497 in the GV pair and 24,212 in the TB pair (Table S3). The maximum-likelihood (ML) tree that included *D. crocata* (2,472 genes; 2,926,723 bases) confirmed the expected phylogenetic relationships (Figure 1), that is, that *D. silvatica* is sister to the two generalist/specialist sister lineages (GV and TB). We estimated that *D. gomerensis* and *D. verneaui* diverged approximately ∿4.1 Mya, whereas the split between *D. tilosensis* and *D. bandamae* occurred ∿3.1 Mya; the age of the common ancestor of these four lineages dates to ∿4.5 Mya (analysis based on 4,539 genes; Figure 1). These estimates are similar to those obtained in Macías-Hernández et al. (2013).

These very recent divergence times, especially the short internal branch lengths, indicated that hemiplasy might represent an important confounding factor in our inferences of convergent evolution. Indeed, although the species tree estimated with ASTRAL had the same fully supported topology (the local posterior support for each branch was 1) than as the ML tree based on the concatenated MSA, the final normalized quartet score of this species tree (0.65) uncover a high gene-tree conflict in our data set. The risk of hemiplasy (HRF) estimated along the species tree obtained with ASTRAL, varied according to the effective population sizes and the examined branch (Figure 3), being small for $N_e \leq 10^4$, high in branches A and C for $N_e \geq 10^5$, and extremely high in all branches for $N_e \geq 10^6$. Given the high fraction of discordant gene trees observed in our data (5,275 out of 7,784 gene trees; 3,666 with high bootstrap support ≥0.75

TABLE 1 Summary of dietary habits, sampling localities, RNA-seq data and assembly statistics for each surveyed *Dysdera* species

	D. silvatica	*D. verneaui*	*D. gomerensis*	*D. bandamae*	*D. tilosensis*
Diet	Generalist	Generalist	Specialist	Generalist	Specialist
Locality (in Canary Island)	La Gomera	Tenerife	El Hierro	Gran Canaria	Gran Canaria
Total raw reads	441,835,864	527,299,202	430,522,240	765,653,462	678,150,384
Total qualified reads	418,205,054	495,937,054	400,095,710	746,925,920	664,654,842
Transcripts	236,283	441,604	213,984	296,544	316,498
Genes (clustered isoforms)	170,846	347,878	177,363	221,801	229,762
Gene average length (in bp)	702	525	622	658	649
Gene maximum length (in bp)	26,709	27,235	27,386	27,369	25,342
HK genes	1,136	1,194	1,232	1,153	1,159
CEG genes	807 (457)	1,180 (457)	1,111 (457)	1,033 (457)	1,143 (457)
GO annotated genes	29,879	38,361	28,158	35,116	37,246
Genes with InterPro domain	30,886	40,771	29,930	37,413	39,480
Functional annotated genes[a]	31,091	41,019	30,106	37,620	39,704
Annotated genes[b]	41,046	51,864	37,087	47,059	50,150
Predicted coding sequences (CDS)	58,966	84,114	55,914	72,352	77,756
% not coding genes	34.51%	24.18%	31.53%	32.62%	33.84%
% not annotated CDS	69.61%	61.66%	66.33%	65.04%	64.50%
1 to 1 orthologs in all species	9,473	9,473	9,473	9,473	9,473
1 to 1 orthologs per species pair	-	19,497	19,497	24,212	24,212

[a]GO or Interpro hits.
[b]GO, Interpro or blast hits.

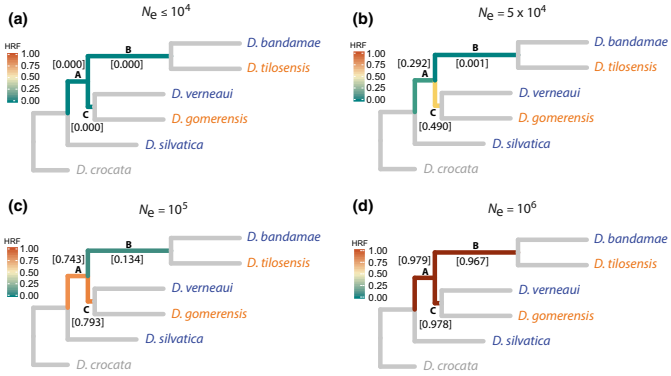


FIGURE 3 Species tree inferred with Astral showing the risk of hemiplasy along the phylogeny. Hemiplasy risk factor values (HRF) were estimated for all internal branches of the tree. The relative probabilities of hemiplasy and homoplasy were inferred under different effective population sizes (N_e; panels a to d) and assuming a fixed mutation rate μ per $2N_e$ generations ($2N_e\mu = 5.5 \times 10^{-3}$). HRF values estimated for all internal branches (in brackets) represent the proportion of discordant traits associated with a branch due to hemiplasy [Colour figure can be viewed at wileyonlinelibrary.com]

in at least one discordant node) together with HRF estimates, the surveyed species (and their ancestors) would have intermediate to high effective population sizes, in a range of $10^4 < N_e \leq 10^6$. Although only a small fraction of these inconsistencies might really affect our inferences of homoplasy (see Section 4), we specifically considered this confounding factor in our study. In contrast, we did not detect the characteristic hallmark of gene flow between extant specialist lineages in the D_{FOIL} analysis of transcripts, neither by analysing all transcripts separately nor by concatenating them in different gene groups (i.e., all transcripts, all candidates, only gene expression or only positive selection candidates; results not shown; see below for the precise definition of each type of candidate).

3.1 | Gene expression changes matching phenotypic convergence: individual gene level

Despite the sex-ratio bias of the studied samples (Table S1), the PCA of the eight *REST* samples of the specialist *D. tilosensis* sequenced separately (four males and four females) showed no evidence of sex-specific expression (Figure S2), which is in agreement with the absence of morphological dimorphism between sexes reported for the Eastern Canarian clade of this genus (Macías-Hernández et al., 2008). We found 774 (out of 19,497) and 1,044 (out of 24,212) genes showing differential expression between specialists and generalist species in the GV and TB pairs, respectively (Figure S3 and Table S4). Remarkably, 147 genes (out of 193) had patterns of gene expression matching phenotypic convergence; that is, the expression profiles had the same trend in both species' pairs with the two specialists significantly under- or overexpressed (hereafter referred to as Matching Gene Expression "MGE" candidates); however, in three cases the tree showed discordant genealogies supported by the entire transcript sequence. The final number of MGE candidates (144 genes) is much higher than that expected by a neutral model of gene expression evolution, both when considering all differentially expressed genes (hypergeometric test; $p = 1.3 \times 10^{-67}$) and separating genes over- or underexpressed in specialist lineages ($p = 2.3 \times 10^{-14}$ and $p = 4.2 \times 10^{-121}$, respectively; hypergeometric test). The proportion of genes significantly underexpressed in specialists was higher both in the two species pairs considered separately (68% in GV and 61% in TB) and, to a much greater extent, across the 144 shared DE candidate genes (114 genes; 79%; Figure 4 and Table S4). All MGE candidates except two functionally uncharacterized proteins (OG9619 and OG15050 in *PALP*) and one phosphatase (OG1641 in *LEGS*) were predominantly expressed in *REST* (Figures 4 and S3), and none of them show DE between males and females of *D. tilosensis* in this body part (results not shown). All these findings indicate that DE analyses are reflecting real differences between specialist and generalist species, and not sex or body part-specific features. Yet, we cannot completely rule out that some of the uncovered candidates were false positives, so they should be considered as promising candidates to be further validated.

Within the biological processes significantly overrepresented (Figure 5a) among MGE candidates, we identified genes involved in the homeostasis of metal ions, catabolism of amino acids, sugars and chitin and activities of enzymes such as phosphatase and hydrolase. The separate analysis according to the direction of gene expression change showed that the 114 MGE candidates downregulated in specialists are significantly enriched in assembly and organization of chromatin, cytoskeleton and other cellular structures (such as the organelles), potential regulation of developmental processes through the smoothened pathway, cell morphogenesis and growth processes, and catabolism of sugars and amino acids. In contrast, the 30 MGE candidates upregulated in specialists are significantly enriched in GO terms associated to the metabolism of steroids, lipids and dicarboxylic acid, the activities of phosphatases and hydrolase, the membrane transport of different substances and responses to various external stimuli including cellular response to oxidative stress. Other interesting but not GO-enriched functions of the MGE candidates include iron ion binding (a predicted cytochrome P450 protein overexpressed in specialist spiders) and zinc ion binding (mostly represented by various putative zinc finger-containing proteins; Table S4). Furthermore, we also found two putative venom toxins among the 144 MGE candidates, one of which encodes a protein similar to the α-latrocrustatoxin (underexpressed in specialists),

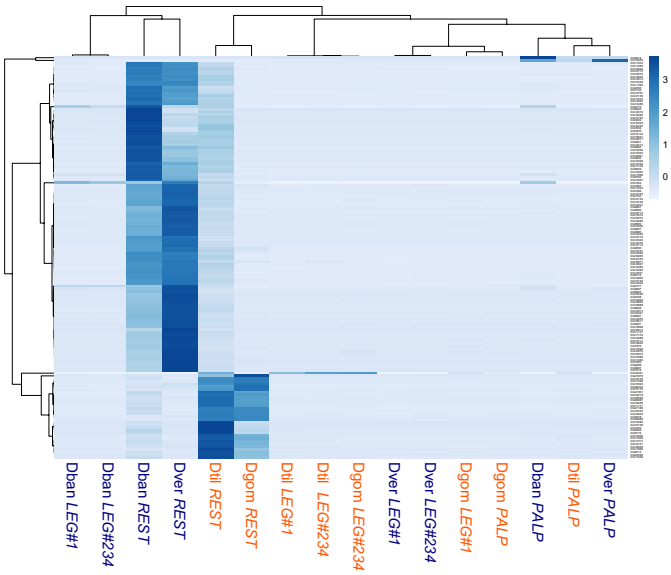


FIGURE 4 Heat map with body part-specific gene expression profiles of the 144 MGE candidates [Colour figure can be viewed at wileyonlinelibrary.com]

while the other is an U32-aranetoxin-Av1a overexpressed in specialists (see Figure S4 and Table S4 for a more detailed functional description of the MGE candidates, including significantly enriched molecular functions).

Our analysis also detected 21 genes specifically expressed in specialists (i.e., with no detectable expression in generalists; referred to as Matching Specialist-specific Expression "MSE" candidates; Figure 2). Fifteen of these *MSE* candidates encode proteins with no significant sequence similarity with any entry in the searched databases; the other six cases, which were not enriched in any GO term, encode catalytic activities, such as hydrolases and peptidases, or are associated with zinc ion-binding proteins, likely involved in the regulation of gene expression (Table S4).

The highly fragmented nature of the transcripts encoding members of the major chemosensory gene families (Vizueta, Rozas, & Sánchez-Gracia, 2018) prevented the credible assignation of many orthogroups and, therefore, a reliable DE analysis comparing specialists and generalists. Besides, for the few orthogroups that could be assigned, we did not find any concordant DE pattern in specialists. The same negative results were obtained for the other orthogroups that showed DE in the chemosensory appendages (*PALP* and *LEG#1* and *LEG#234*) in the study of Vizueta et al. (2017).

3.2 | Gene expression changes matching phenotypic convergence: gene function level

Apart from the 144 MGE candidates, the group of genes with DE only in one species pair, 627 in GV pair and 897 in TB pair, respectively, also shared a significant number of enriched GO terms (70 terms; hypergeometric test, $p = 4.7 \times 10^{-11}$ for all DE genes; $p = 2.2 \times 10^{-23}$ and $p = 1.3 \times 10^{-2}$ for under- and overexpressed genes, respectively). Remarkably, some of these GO terms are the same as those overrepresented among the MGE candidates. For the genes underexpressed in specialists, these included chromatin assembly, the organization of cellular components, such as the cytoskeleton or organelles, and cell growth. Other additional functions, such as phosphate metabolism regulation and the apoptotic process involved in morphogenesis, are also shared among these genes. For the genes overexpressed in specialists, the enriched functions shared between species pairs include lipid catabolism, oxidation-reduction process and response to antibiotics (Figure S4 and Table S4).

Among the orthogroups with DE only in one species pair but with equivalent functions, we found genes involved in detoxification processes and genes encoding various members of the cytochrome P450 family (most of them overexpressed in specialists, seven and nine different copies in the GV and TB pairs, respectively) or proteins with esterase activity (seven and six of these enzymes in the GV and TB pairs, respectively). Additionally, we found 29 putative venom toxin-encoding genes in the GV pair (eight overexpressed in G) and 34 in the TB pair (26 overexpressed in T). Interestingly, although the encoding genes differed between the two specialists, they had very similar predicted functions, such as astacin-like metalloprotease toxin precursors or aranetoxin-Av1a and latrotoxins, among others (Table S4).

3.3 | Positive selection matching phenotypic convergence: individual gene level

We applied the aBSREL model to estimate the distribution of ω values of all single-copy orthologus with complete sequences and

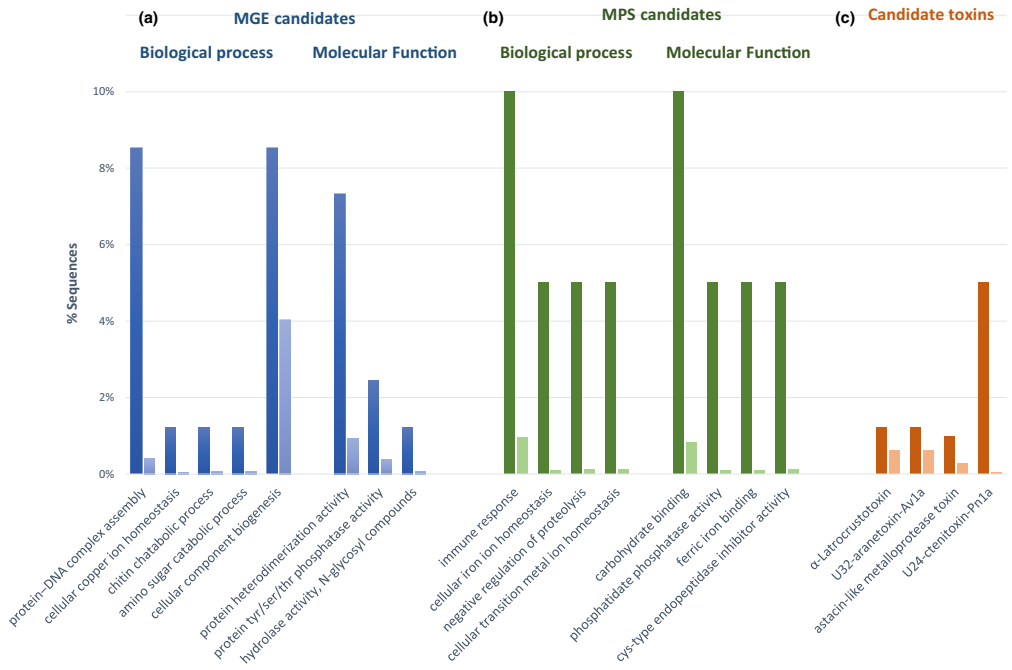


FIGURE 5 Bar charts with the most relevant results of the GO enrichment analyses (see Figure S3 for more detailed versions). (a) Orthogroups with differential expression profiles matching phenotypic convergence (144 MGE candidates); (b) Orthogroups under positive selection in the two specialists (17 MPS candidates); (c) Most representative candidates encoding venom toxins in stenophagous *Dysdera*. Dark and light tones represent the proportion of genes with a given associated GO in the candidate and the population (whole transcriptome) set, respectively [Colour figure can be viewed at wileyonlinelibrary.com]

without premature stop codons (7,784 genes; Figure 2 and Table S3). This genome-wide analysis uncovered opposite trends between GV and TB pairs; while the overall selective constraints appear to have been relaxed in the *D. tilosensis* lineage, they intensified in the *D. gomerensis* branch (Figure S5). Nevertheless, the analysis of individual genes identified nine genes with significant differences in the selective constraint values shared between the two specialists (or the two generalists; RELAX framework analysis, FDR of 0.2; Table S5; referred as Matching Functional Constraint "MFC" candidates). Six of these candidates showed the relaxation hallmark in specialists, while the other three showed a significant increase in the selective constraint. We found some overrepresented biological functions among MFC candidates, such as carbohydrate metabolism and homeostasis, neuropeptide signalling, tRNA modification and pyridine metabolism (Figure S4). When we considered not enriched GO terms, the genes with increased functional constrains in specialists encode proteins similar to the membrane glycoprotein LIG-1, a neuropeptide receptor-like protein and zinc finger proteins, while the genes that have relaxed most in specialist's species encode two zinc finger-like proteins and a hexokinase.

We identified 297 genes with significant evidence of positive selection in specialist lineages, 169 in *D. gomerensis*, 150 in *D. tilosensis* and, remarkably, 22 cases in which positive selection was inferred in both dietary specialists (Figure 2 and Table S6; referred to as Matching Positive Selections "MPS" candidates). After excluding five coding regions with discordant genealogies supported by the entire transcript sequence, the number of MPS candidates (17) is clearly greater than that expected by chance (across the 297 genes showing positive selection in specialists; hypergeometric test; $p = 1.5 \times 10^{-8}$). These genes are enriched in biological processes such as germ cell migration and cell death, cell junction assembly and organization, regulation of the immune response or iron ion homeostasis (Figures 5 and Figure S4). Interestingly, one of these genes with endopeptidase inhibitor activity encodes a protein with sequence similarity to U24-ctenitoxin-Pn1a, a possible venom toxin related to cysteine proteinase inhibitors.

The PCOC method (Rey et al., 2018) identified convergent shifts in amino acid preferences in 14 out of the 17 MPS candidates (FDR = 0.03%; TPR = 99.7%; Figure 6; Table S6 and Figure

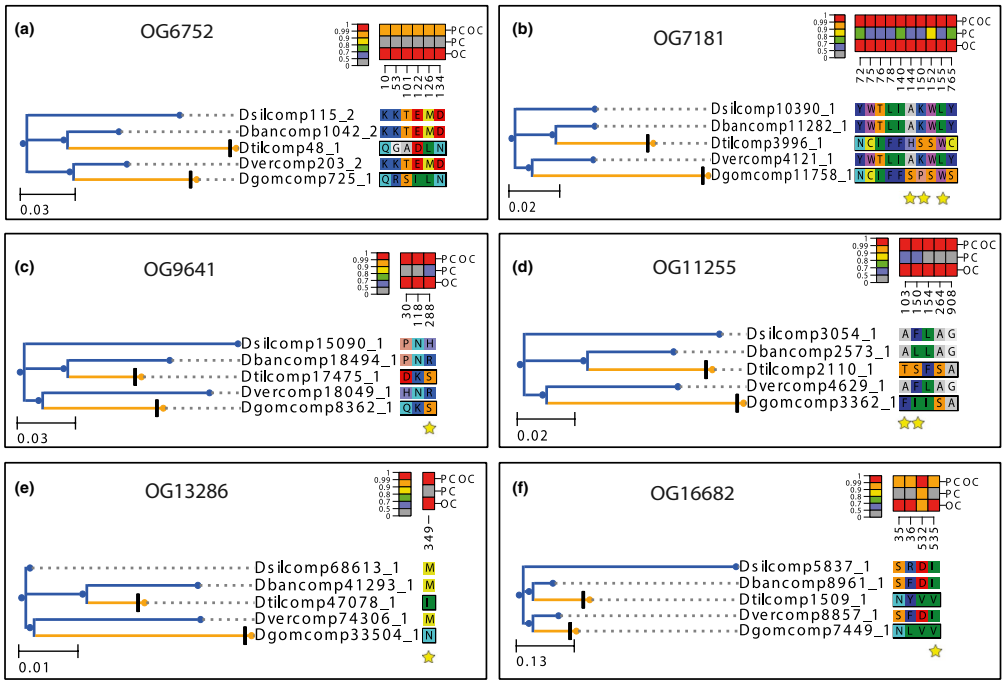


FIGURE 6 Relevant orthogroups showing evidence of convergent amino acid substitutions. (a) Orthogroup encoding the venom toxin OG6752. (b–f) Orthogroups with positions evolving under positive selection. Amino acid positions are shaded with different tones according to their profiles, and only positions with a *PP* equal to or greater than 0.99 according to the PCOC, PC or OC model are shown (Rey et al., 2018). Stars highlight the sites identified as being positively selected in MEME [Colour figure can be viewed at wileyonlinelibrary.com]

S6). Furthermore, in five cases, the subsequent MEME analysis indicated that some of the amino acid sites involved in these convergent shifts have also evolved by positive selection (eight amino acid sites; Figure 6). The target genes include (a) the U24-ctenitoxin-Pn1a candidate toxin (OG6752 orthogroup; six amino acid changes); (b) OG7181, a transcript encoding a protein similar to tectonin (10 amino acid changes, 3 of them under); (c) OG9641, a transcript encoding a protein involved in response to oxidative stress (three amino acid changes, one of them also detected with MEME); (d) OG11255, a gene that encodes a product similar to a mannose receptor (five amino acid changes, two of them also detected with MEME); (e) OG13286, a protein likely encoding a sodium channel (one amino acid change, also detected with MEME); and (f) OG16682, a hydrolase involved in nitrogen compound metabolism (four amino acid changes, one of them detected with MEME). The analysis also inferred some amino acid substitutions responsible of a convergent shift of preferences in specialists but without evidence of positive selection in OG9529, a putative dehydrogenase and oxidoreductase (four amino acids; Figure S6).

3.4 | Positive selection matching phenotypic convergence: gene function level

Although the group of genes under positive selection in only one of the two specialists (147 in GV pair and 138 in TB pair, respectively) did not share more significantly enriched GO terms than expected by change (only three shared GO were enriched in both pairs; hypergeometric test; p = .19), the number of total GO terms shared by these two groups is greater than expected ($p = 5.3 \times 10^{-75}$ based on the hypergeometric distribution). Among shared GO terms, we found processes and functions such as chitin metabolism (including proteolysis activity), lipid metabolism, metal ion binding (zinc in both pairs, copper in *D. gomerensis* and iron in *D. tilosensis*), and hydrolase and oxidoreductase activities (Figure S4). In addition, we also detected the signature of positive selection in six genes encoding putative venom toxins: four in *D. gomerensis* and two in *D. tilosensis* (Table S6).

The gene family analysis also uncovered the hallmark of positive selection in five gene families affecting both specialist lineages (Figure 2 and Table S6). One family (the OG3133 orthologous group),

which included sequences without any functional annotation, also showed copy number variation in the two specialists (two and three copies in *D. gomerensis* and *D. tilosensis*, respectively, compared to one in the generalist species). The other four gene families encoded proteins with possible functions in chitin metabolism and sequences similar to carbohydrate and zinc ion-binding proteins, hydrolases and other enzymes with catalytic activity. Again, we found a gene family encoding putative venom components (in this case, with no characterized target) among positively selected gene families.

4 | DISCUSSION

The evolution of stenophagy, dietary specialization from a generalist ancestor, most likely involves gene regulatory changes, amino acid replacements in proteins and/or even copy number variation in gene families. Here, we focused our analysis on the first two issues since comparative transcriptomics based on de novo assemblies prevents accurate estimation of changes in gene expression and gains and losses in gene family members. Our approach allows detecting genetic changes in the genes expressed in adults (either in the same gene or in equivalent gene functions) matching the phenotypic convergence observed in dietary specialist *Dysdera*. Nevertheless, it is largely known that hemiplasy can also produce such matching patterns, inducing false evidence of convergent evolution (Mendes et al., 2016; Wu et al., 2018). Indeed, the high level of gene-tree discordance caused by ancestral polymorphisms could potentially explain some of the repeated changes identified in *D. gomerensis* and *D. tilosensis*. Nonetheless, some lines of evidence support that most of the candidates reported in this study accumulated convergent changes in specialist lineages. First, for realistic effective population sizes (i.e., $10^4 < N_e \leq 10^5$; these spiders are island endemic predators with likely low census sizes), the probability of observing discordant trees matching the phenotypic convergence is very low (Figure 3). The estimates of the HRF values in branch B under realistic effective population sizes ranged from 0.001 to 0.134 (Figure 3b,c). Therefore, the probability of occurrence of ILS on this branch, accompanied by a mutation in the branch A or in an older lineage creating a false pattern of homoplasy, is much lower than that of true homoplasy (Guerrero & Hahn, 2018). Second, among the total set of discordant gene trees with high bootstrap support, only the 1.69% (62 out of 3,666) yielded resolved topologies that match exactly the one expected from convergence in specialists, which agrees with hemiplasy risk predictions for intermediate effective population sizes. Even so, and to be conservative, we excluded from the downstream functional prediction analysis all candidates with gene trees included in this 1.69%. This approach, however, may not be suitable for detecting convergent changes in gene expression in specialists. Actually, the assumption that the regulatory regions responsible of the concordant changes in gene expression of candidate genes are completely linked to the transcribed sequence (i.e., both share the same gene tree) may not be correct. Estimates of the recombination rate in these genomes are not available and, more importantly, some of these mutations could be far away from the coding region, even acting in *trans*. In these cases, however, we would expect that gene-tree discordance will be randomly distributed across the genome. We found, by contrast, a clear bias in our candidates towards genes and functions biologically relevant for dietary specialists. Bearing all this in mind, the fixation of convergent genetic changes remains as the most likely explanation for most of the discordant patterns matching phenotypic convergence, even for MGE candidates. Consequently, we demonstrated that our study design, with two evolutionary replicates of the same dietary specialization event, was able to identify potential candidate genes and groups of functionally equivalent genes responsible in part to these remarkable ecological shifts.

A priori, we would expect that the biological functions targeted by selection are related to prey capture and food assimilation, both in digestive and metabolic aspects. Since genetic changes underlying morphological modifications of the specialists' mouthparts likely involve changes in gene expression patterns during development, they were undetectable in our comparative analysis of adult transcriptomes. However, other aspects related to the detection, attack, consumption and digestion of a prey with remarkable behavioural and chemical defences definitely played a crucial role in specialization. Several studies have revealed significant differences in the growth and nutrient extraction efficiencies in specialist *Dysdera* fed on woodlouse, which suggests the existence of metabolic adaptations (Řezáč & Pekár, 2007; Toft & Macías-Hernández, 2017; Macías-Hernández et al., *in preparation*). Toxicity is the most relevant nutritional aspect that makes isopods a prey commonly rejected by most generalist spiders (Hopkin & Martin, 1985). Indeed, isopods accumulate toxic substances, including high concentrations of heavy metals from the soil, especially copper but also zinc, lead and cadmium, in vesicles such as lysosomes (Paoletti & Hassall, 1999). The toxic effects as well as some of the underlying genetic response mechanisms of heavy metals on terrestrial invertebrates have been known for a long time (Janssens, Roelofs, & van Straalen, 2009; Merritt & Bewick, 2017; Migula, Wilczek, & Babczyńska, 2013). Remarkably, our results are in full agreement with the few comparative transcriptomics studies conducted on these types of animals under different metal stress conditions (e.g,. Gomes, Scott-Fordsmand, & Amorim, 2014; Roelofs et al., 2009; Zapata, Tanguy, David, Moraga, & Riquelme, 2009), including in spiders (Li et al., 2016). These studies demonstrate that arthropods exposed to heavy metals show important gene expression changes relative to controls; remarkably, some of the reported gene targets also appear among our MGE candidates or correspond with some of the molecular functions enriched in our list. Some examples include ABC transporters, amiloride-sensitive sodium channels, ATPases, MAP kinases, ubiquitin ligases, histones, members of the cytochrome P450 family and ribosomal proteins (Table S4). These consistent results across different studies on phylogenetically distant species support the idea of a relatively well-conserved common mechanism for the

tolerance of heavy metal toxicity across animals. The old origin of such an evolutionary mechanism validates our approach for identifying the genetic determinants of stenophagy in *Dysdera*.

4.1 | Genetic changes matching phenotypic convergence: metal-induced damage or adaptive response to metal stress?

We found that most MGE candidates were specifically downregulated in specialists and encoded molecular functions involved in cell response, vesicular transport, organization of organelles and cytoskeleton, cilia assembly or cell adhesion (Table S4). Noticeably, these are the most frequent cell modifications observed in intestinal tissue damage by heavy metals from the diet (e.g., Bednarska et al., 2016; Köhler & Alberti, 1992; Zhang et al., 2001). Indeed, in soil arthropods subjected to heavy metal stress, midgut cells show evident histological modifications indicative of metal deposition in intracellular granules and gut epithelial degeneration. Although the downregulation pattern observed in specialist *Dysdera* could be the result of a direct stress-induced perturbation of gene expression caused by the high concentration of heavy metals supplied in a woodlouse-rich diet, they might actually be part of an adaptive biological response to excrete metals or other toxic substances more efficiently, thus avoiding their assimilation (Van Straalen & Roelofs, 2005). Consistent with this hypothesis, we observed concordant DE patterns in some MAP kinase pathway members, which participate in an important stress-activated/immune response cascade (Chmielowska-Bąk & Deckert, 2012), and in some ubiquitin ligases, which, among other functions, are involved in the inhibition of cell growth and cycle arrest in response to DNA damage (Cao & Yan, 2012). The adaptive response in specialists would consist of downregulating a set of genes to keep gut epithelial cells in a semi-degenerated functional and structural state that allows enhanced accumulation of heavy metals in granules and very fast and effective intestinal exfoliation and regeneration.

Our analysis also uncovered a number of upregulated MGE and MPS candidates associated with iron, copper and zinc binding and homeostasis, which can also be part of an adaptive mechanism of detoxification in specialist *Dysdera*. Among these candidates, we found amiloride-sensitive sodium channels, membrane ATPases and ABC and dicarboxylate transporters. These proteins are either antiporters for metal cations or are involved in cellular mechanisms for heavy metal vacuolar sequestration (Ahearn, Sterling, Mandal, & Roggenbeck, 2010) or in cellular metal homeostasis and detoxification (e.g., Lee, Yang, Zhitnitsky, Lewinson, & Rees, 2014; Sooksa-Nguan et al., 2009). Another set of interesting candidates are the proteins annotated as syntaxin-5-like proteins with a SNARE domain, which are involved in vesicle tethering and fusion associated with copper ion homeostasis (Norgate et al., 2010) and, in addition to being significantly overexpressed in both specialists, also show signals of positive selection in *D. tilosensis*.

It is well known that heavy metal-associated toxicity is largely due to damage to the oxidative tissue caused by the accumulation of reactive oxygen species in the cell (Schieber & Chandel, 2014). Noticeably, among the upregulated MGE candidates (and those regulated in only one of the specialists), we found members of family 3 of the P450 cytochromes, a group of monooxygenases that constitute the largest and most functionally diverse class of insect detoxification enzymes and that have been implicated in the oxidative detoxification of furanocoumarins, alkaloids, plant secondary metabolites and synthetic insecticides (Nelson & Nebert, 2011). Additionally, we identified among the candidates several esterases, a group of proteins with a role in heavy metal and pesticide detoxification that have been used as biomarkers of metal exposure in many organisms, including spiders (Wilczek, Babczyńska, Migula, & Wencelis, 2003). We identified esterases significantly overexpressed in both specialists, although in this case, the orthogroups of *D. gomerensis* and *D. tilosensis* were different, suggesting possible convergence at the functional level rather than at the gene level. Remarkably, two of these esterases also showed a positive selection signal in *D. gomerensis*.

We also detected other MGE candidates associated with the metabolism of some essential nutrients, such as proteins with chitin-binding and chitinase activity, and enzymes involved in the metabolism of amino acids, sugars and lipids. Given that most of these candidates were downregulated in specialists, the adaptive advantage could be associated with a reduction in biosynthetic processes to save energy, presumably to dedicate the energy to detoxification processes. However, the presence of some upregulated and positively selected genes among these metabolic candidates indicates that specialists might also have developed an adaptive mechanism to enhance the assimilation and metabolization of some other nutrients present in woodlice but less accessible to other preys.

Finally, it is worth noting that MPS candidates are also significantly enriched in genes related to the immune system. It has been reported that high concentrations of heavy metals negatively affect important processes, such as phagocytosis and chemotaxis, during the generation of the immune response (Boyd, 2010). The footprint of positive selection detected in specialist *Dysdera*, matching phenotypic divergence, might reflect an adaptive mechanism to alleviate the negative immunomodulation effects of heavy metals. In fact, there is evidence that positive selection promoted local adaptation of herbivore insects to heavy metal polluted environments by enhancing immune functions (van Ooik & Rantala, 2010), suggesting the important adaptive character of this system under metal stress conditions.

4.2 | A possible role of venom toxins in the convergent dietary shift

Stenophagous spiders (e.g., myrmecophagous, termitophagous and araneophagous spiders) show increased venom toxicity to the preferred prey, while related generalists show similar toxicities to all preys (Pekár, Líznarová, Bočánek, & Zdráhal, 2018). The analysis of venom components in stenophagous species indicates that this difference in efficacy is caused by the presence of prey-specific toxins, suggesting evolutionary adaptations for more effective exploitation of focal prey. Notably, we identified a number of transcripts encoding

venom toxins among the MGE candidates, most of which were up-regulated in specialists, an opposite pattern to that obtained for the rest of the MGE candidates. Among others, we found candidates encoding astacin-like metalloproteases. Astacins share common features with serralysins, matrix metallo-endopeptidases and snake venom proteases and might be involved in the proteolytic processing of other venom toxins or even play a role in extra-oral digestion of prey, which could be important in the specialization of Canarian *Dysdera* to woodlice. Interestingly, the MGE candidates encoding astacin-like metalloproteases belonged to different orthogroups in each specialist species, which suggests an additional example of functional convergence through different genes. Our analysis also uncovered other candidates that encode some lesser-known toxins, such as products with sequence similarity to U24-ctenitoxin-Pn1a (presumably a protease inhibitor), pisautoxin-Dm1a (a toxin from the venom of the spider *Dolomedes mizhoanus* with an unknown target), alpha-latrotoxins (which induce massive neurotransmitter release) and aranetoxins (also with an unknown target). Remarkably, we found that among the alpha-latrotoxins, a transcript with similarity to a crustacean-selective component of spider venom (the alpha-latrocrustatoxin; Grishin, 1998) also showed the signature of positive selection, making it a promising candidate for stenophagy. Further research including venom gland-specific transcriptomes and the study of venom toxicity to different preys would be required to shed light on the role of venom in the convergent dietary specialization of *Dysdera*.

4.3 | Repeated adaptation to stenophagy in Canarian endemic *Dysdera*: collateral or parallel evolution?

Here, we uncovered several pieces of evidence supporting the adaptive divergence hypothesis in stenophagous *Dysdera* inhabiting Western Canary Islands. First, the functional annotation of the majority of genes with concordant changes in gene expression between generalist and specialist spiders clearly points towards an active role of these genes in the dietary shift. Second, we detected repeated episodes of positive selection in the same genes (or functionally related group of genes) in the two specialists' lineages. Furthermore, a significant number of MPS candidates showed convergent amino acid preference shifts in the two focal branches, some of which were also inferred to be under positive selection. Altogether, these results provide new significant evidence that species can find the same molecular solutions to adapt predictably to similar ecological niches more often than previously thought (see Marques et al., 2017; Nosil et al., 2018, for other recent examples).

Specialist *Dysdera* may have repeatedly adapted to stenophagy through parallel or collateral evolution. In the first case, convergence would result from the accumulation of the same or similar mutations in evolutionary independent lineages, whereas in the second, selection on either shared ancestral or introgressed variations would be the responsible of the convergent patterns (Stern, 2013). In recent years, increasing evidence has emerged suggesting the important role of shared genetic variation as a substrate for driving repeated evolution of ecotypes in nature (e.g., Jones et al., 2012; Marques, Meier, & Seehausen, 2019; Schluter & Conte, 2009; Van Belleghem et al., 2018). Our genome-wide HRF and D_{FOIL} analyses point to that most of our candidates originated from parallel independent evolution (i.e., relatively low risk of random ILS and nonsignificant D_{FOIL} results). On the other hand, in the five positive selection candidates where the individual gene trees were incongruent, the apparent homoplasy could be the result of collateral evolution. Unfortunately, in these cases, current data would not allow to disentangle collateral evolution from random ILS at the individual gene level. Accordingly, and to avoid reporting candidates with false patterns of homoplasy, we excluded these five genes with discordant topologies, restricting the analysis on the parallel fixation of de novo mutations. Further research including polymorphism from whole genome data would be needed to unequivocally establish the relative role of collateral evolution in the convergence observed in these island endemic spiders.

Altogether, our findings suggest that the ecological opportunity provided by the colonization of the Canary Islands facilitated the exploration of multiple adaptive landscapes by *Dysdera* and its diversification on similar peaks (Mahler, Ingram, Revell, & Losos, 2013), providing an exceptional example of repeatability in evolution and shedding light on the genetic determinants of phenotypic convergence (Stroud & Losos, 2016). Besides, our results support the idea that convergence can involve repeated changes at different hierarchical levels (Rosenblum, Parent, & Brandt, 2014). We found convergent changes at the amino acid, gene and gene function levels that would be mostly associated to the excretion and detoxification of heavy metals accumulated in the preferred prey, and some venom components likely related to prey capture. We also demonstrated that natural selection promoted the fixation of some of these changes, confirming the view that adaptive forces are a primary determinant of phenotypic convergence (Storz, 2016). Moreover, our report uncovering repeated genetic changes in pairs of phylogenetically close taxa supports the ongoing debate that the probability of shared molecular changes for convergent phenotypes correlates with node age (Conte, Arnegard, Peichel, & Schluter, 2012). Hence, this study not only provides new evidence on the genomic basis of an extraordinary example of a convergent ecological shift in a non-model organism but also offers new insights into the long-standing debate about predictability in evolution.

ACKNOWLEDGEMENTS

We thank to five anonymous reviewers for their useful comments on the manuscript. We also thank Cristina Frías-López for helping with the RNA extractions, and Matthew Hahn for his suggestions and recommendations. This work was supported by the Ministerio de Economía y Competitividad of Spain (CGL2012-36863, CGL2013-45211, CGL2016-75255 and CGL2016-80651) and the Comissió Interdepartamental de Recerca I Innovació Tecnològica of Catalonia, Spain (2014SGR1055 and 2014SGR1604). J.V. was supported by a FPI grant (Ministerio de Economía y Competitividad of Spain, BES-2014-068437). We acknowledge the Cabildos of

Tenerife, Gran Canaria and La Gomera, as well as the Garajonay National Park that have granted us collection permits, and often also helped with lodging and logistics during campaigns.

CONFLICT OF INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

A.S-G. and J.R. designed, conceived and supervised the research; N.M-H and M.A.A. provided the biological material. J.V. performed the experiments and the bioinformatics work, and analysed the data. M.A.A. performed the dissecting analysis and participated in the data interpretation. J.V., J.R. and A.S-G. wrote the first version of the manuscript. N.M-H. and M.A.A. revised the manuscript and participated in the writing of the final version. All authors read and approved the final version of the manuscript.

DATA AVAILABILITY STATEMENT

The raw sequence data generated for this work have been deposited at the Sequence Read Archive (SRA) under Bioproject PRJNA437566. Additional data and analysis generated in this study have been deposited in Figshare (https://doi.org/10.6084/m9.figshare.7726508.v1).

ORCID

Joel Vizueta https://orcid.org/0000-0003-0139-3013
Nuria Macías-Hernández https://orcid.org/0000-0003-4759-3619
Miquel A. Arnedo https://orcid.org/0000-0003-1402-4727
Julio Rozas https://orcid.org/0000-0002-6839-9148
Alejandro Sánchez-Gracia https://orcid.org/0000-0003-4543-4577

REFERENCES

Ahearn, G. A., Sterling, K. M., Mandal, P. K., & Roggenbeck, B. (2010). *Heavy metal transport and detoxification by crustacean epithelial lysosomes*. *Epithelial transport physiology*. Totowa, NJ: Humana Press.

Almén, M. S., Lamichhaney, S., Berglund, J., Grant, B. R., Grant, P. R., Webster, M. T., & Andersson, L. (2016). Adaptive radiation of Darwin's finches revisited using whole genome sequencing. *BioEssays*, *38*(1), 14–20. https://doi.org/10.1002/bies.201500079

Arnedo, M. (2001). Radiation of the spider genus *Dysdera* (Araneae, Dysderidae) in the Canary Islands: Cladistic assessment based on multiple data sets. *Cladistics*, *17*(4), 313–353. https://doi.org/10.1006/clad.2001.0168

Arnedo, M. A., Oromí, P., Múrria, C., Macías-Hernández, N., & Ribera, C. (2007). The dark side of an island radiation: Systematics and evolution of troglobitic spiders of the genus *Dysdera* Latreille (Araneae:Dysderidae) in the Canary Islands. *Invertebrate Systematics*, *21*(6), 623–660. https://doi.org/10.1071/IS07015

Avise, J. C., & Robinson, T. J. (2008). Hemiplasy: A new term in the lexicon of phylogenetics. *Systematic Biology*, *57*(3), 503–507. https://doi.org/10.1080/10635150802164587

Bednarska, A. J., Laskowski, R., Pyza, E., Semik, D., Świątek, Z., & Woźnicka, O. (2016). Metal toxicokinetics and metal-driven damage to the gut of the ground beetle *Pterostichus oblongopunctatus*. *Environmental Science and Pollution Research*, *23*(21), 22047–22058. https://doi.org/10.1007/s11356-016-7412-8

Benjamini, Y. H., & Hochberg, Y. (1995). Controlling the false discovery rate – A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*, 289–300. https://doi.org/10.2307/2346101

Boyd, R. S. (2010). Heavy metal pollutants and chemical ecology: Exploring new Frontiers. *Journal of Chemical Ecology*, *36*(1), 46–58. https://doi.org/10.1007/s10886-009-9730-5

Cao, J., & Yan, Q. (2012). Histone ubiquitination and deubiquitination in transcription, DNA damage response, and cancer. *Frontiers in Oncology*, *2*, 26. https://doi.org/10.3389/fonc.2012.00026

Chmielowska-Bąk, J., & Deckert, J. (2012). A common response to common danger? Comparison of animal and plant signaling pathways involved in cadmium sensing. *Journal of Cell Communication and Signaling*, *6*(4), 191–204. https://doi.org/10.1007/s12079-012-0173-3

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). BLAST2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, *21*(18), 3674–3676. https://doi.org/10.1093/bioinformatics/bti610

Conte, G. L., Arnegard, M. E., Peichel, C. L., & Schluter, D. (2012). The probability of genetic parallelism and convergence in natural populations. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1749), 5039–5047. https://doi.org/10.1098/rspb.2012.2146

Coyne, J. A., & Orr, H. A. (2004). *Speciation*. Sunderland, MA: Sinauer Associates.

Dejean, A. (1997). Distribution of colonies and prey specialization in the ponerine ant genus *Leptogenys* (Hymenoptera: Formicidae). *Sociobiology*, *29*, 293–299.

Drobne, D. (1997). Terrestrial isopods-a good choice for toxicity testing of pollutants in the terrestrial environment. *Environmental Toxicology and Chemistry*, *16*(6), 1159–1164. https://doi.org/10.1002/etc.5620160610

Emerson, B. C. (2002). Evolution on oceanic islands: Molecular phylogenetic approaches to understanding pattern and process. *Molecular Ecology*, *11*(6), 951–966. https://doi.org/10.1046/j.1365-294X.2002.01507.x

Falcon, S., & Gentleman, R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics*, *23*(2), 257–258. https://doi.org/10.1093/bioinformatics/btl567

Fernández, R., Hormiga, G., & Giribet, G. (2014). Phylogenomic analysis of spiders reveals nonmonophyly of Orb weavers. *Current Biology*, *24*(15), 1772–1777. https://doi.org/10.1016/j.cub.2014.06.035

Frías-López, C., Almeida, F. C., Guirao-Rico, S., Vizueta, J., Sánchez-Gracia, A., Arnedo, M. A., & Rozas, J. (2015). Comparative analysis of tissue-specific transcriptomes in the funnel-web spider *Macrothele calpeiana* (Araneae, Hexathelidae). *PeerJ*, *3*, e1064. https://doi.org/10.7717/peerj.1064

Gillespie, R. (2004). Community assembly through adaptive radiation in Hawaiian spiders. *Science*, *303*(5656), 356–359. https://doi.org/10.1126/science.1091875

Gillespie, R. G., & Roderick, G. K. (2002). Arthropods on islands: Colonization, speciation, and conservation. *Annual Review of Entomology*, *47*(1), 595–632. https://doi.org/10.1146/annurev.ento.47.091201.145244

Gomes, S. I. L., Scott-Fordsmand, J. J., & Amorim, M. J. B. (2014). Profiling transcriptomic response of *Enchytraeus albidus* to Cu and Ni: Comparison with Cd and Zn. *Environmental Pollution*, *186*, 75–82. https://doi.org/10.1016/j.envpol.2013.11.031

Gorvett, H. (1956). Tegumental glands and terrestrial life in woodlice. *Proceedings of the Zoological Society of London*, *126*(2), 291–314. https://doi.org/10.1111/j.1096-3642.1956.tb00439.x

Grant, P. R., & Grant, B. R. (2008). *How and why species multiply: The radiation of Darwin's finches*. Princeton, NJ: Princeton University Press.

Grishin, E. V. (1998). Black widow spider toxins: The present and the future. *Toxicon*, *36*(11), 1693–1701.

Guerrero, R. F., & Hahn, M. W. (2018). Quantifying the risk of hemiplasy in phylogenetic inference. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(50), 12787–12792. https://doi.org/10.1073/pnas.1811268115

Henning, F., & Meyer, A. (2014). The evolutionary genomics of cichlid fishes: Explosive speciation and adaptation in the postgenomic era. *Annual Review of Genomics and Human Genetics*, *15*(1), 417–441. https://doi.org/10.1146/annurev-genom-090413-025412

Hopkin, S. P., & Martin, M. H. (1985). Assimilation of zinc, cadmium, lead, copper, and iron by the spider *Dysdera crocata*, a predator of woodlice. *Bulletin of Environmental Contamination and Toxicology*, *34*(1), 183–187. https://doi.org/10.1007/BF01609722

Janssens, T. K. S., Roelofs, D., & van Straalen, N. M. (2009). Molecular mechanisms of heavy metal tolerance and evolution in invertebrates. *Insect Science*, *16*(1), 3–18. https://doi.org/10.1111/j.1744-7917.2009.00249.x

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., … Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*(7392), 55–61. https://doi.org/10.1038/nature10944

Juan, C., Emerson, B. C., Oromí, P., & Hewitt, G. M. (2000). Colonization and diversification: Towards a phylogeographic synthesis for the Canary Islands. *Trends in Ecology & Evolution*, *15*(3), 104–109. https://doi.org/10.1016/S0169-5347(99)01776-0

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), 27–30. https://doi.org/10.1093/nar/28.1.27

Köhler, H.-R., & Alberti, G. (1992). The effect of heavy metal stress on the intestine of diplopods. *Berichte Naturwissenschaftlich-Medizinischer Verein Innsbruck*, *10*, 257–267.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with BOWTIE 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Lee, J. Y., Yang, J. G., Zhitnitsky, D., Lewinson, O., & Rees, D. C. (2014). Structural basis for heavy metal detoxification by an Atm1-type ABC exporter. *Science*, *343*(6175), 1133–1136. https://doi.org/10.1126/science.1246489

Li, B., & Dewey, C. N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*, 323. https://doi.org/10.1186/1471-2105-12-323

Li, C.-C., Wang, Y., Li, G.-Y., Yun, Y.-L., Dai, Y.-J., Chen, J., & Peng, Y. (2016). Transcriptome profiling analysis of wolf spider *Pardosa pseudoannulata* (Araneae: Lycosidae) after cadmium exposure. *International Journal of Molecular Sciences*, *17*(12), 2033. https://doi.org/10.3390/ijms17122033

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, *13*(9), 2178–2189. https://doi.org/10.1101/gr.1224503

Losos, J. B., Arnold, S. J., Bejerano, G., Brodie, E. D., Hibbett, D., Hoekstra, H. E., … Turner, T. L. (2013). Evolutionary biology for the 21st Century. *PLoS Biology*, *11*(1), e1001466. https://doi.org/10.1371/journal.pbio.1001466

Losos, J. B., Jackman, T. R., Larson, A., Queiroz, K., & Rodriguez-Schettino, L. (1998). Contingency and determinism in replicated adaptive radiations of island lizards. *Science*, *279*(5359), 2115–2118.

Losos, J. B., & Ricklefs, R. E. (2009). Adaptation and diversification on islands. *Nature*, *457*(7231), 830–836. https://doi.org/10.1038/nature07893

MacArthur, R. H., & Wilson, E. O. (1967). *The theory of island biogeography*. Princeton, NJ: Princeton University Press.

Machado, A., Rodríguez-Expósito, E., López, M., & Hernández, M. (2017). Phylogenetic analysis of the genus *Laparocerus*, with comments on colonisation and diversification in Macaronesia (Coleoptera, Curculionidae, Entiminae). *ZooKeys*, *651*, 1–77. https://doi.org/10.3897/zookeys.651.10097

Macías-Hernández, N., Bidegaray-Batista, L., Emerson, B. C., Oromí, P., & Arnedo, M. A. (2013). The imprint of geologic history on within-island diversification of woodlouse-hunter spiders (Araneae, Dysderidae) in the Canary Islands. *The Journal of Heredity*, *104*(3), 341–356. https://doi.org/10.1093/jhered/est008

Macías-Hernández, N., de la Cruz López, S., Roca-Cusachs, M., Oromí, P., & Arnedo, M. A. (2016). A geographical distribution database of the genus *Dysdera* in the Canary Islands (Araneae, Dysderidae). *ZooKeys*, *625*, 11–23. https://doi.org/10.3897/zookeys.625.9847

Macías-Hernández, N., Oromí, P., & Arnedo, M. A. (2008). Patterns of diversification on old volcanic islands as revealed by the woodlouse-hunter spider genus *Dysdera* (Araneae, Dysderidae) in the eastern Canary Islands. *Biological Journal of the Linnean Society*, *94*(3), 589–615. https://doi.org/10.1111/j.1095-8312.2008.01007.x

Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, *46*(3), 523–536. https://doi.org/10.1093/sysbio/46.3.523

Mahler, D. L., Ingram, T., Revell, L. J., & Losos, J. B. (2013). Exceptional convergence on the macroevolutionary landscape in island lizard radiations. *Science*, *341*(6143), 292–295. https://doi.org/10.1126/science.1232392

Marques, D. A., Meier, J. I., & Seehausen, O. (2019). A combinatorial view on speciation and adaptive radiation. *Trends in Ecology & Evolution*, *34*(6), 531–544. https://doi.org/10.1016/j.tree.2019.02.008

Marques, D. A., Taylor, J. S., Jones, F. C., Di Palma, F., Kingsley, D. M., & Reimchen, T. E. (2017). Convergent evolution of SWS2 opsin facilitates adaptive radiation of threespine stickleback into different light environments. *PLOS Biology*, *15*(4), e2001627. https://doi.org/10.1371/journal.pbio.2001627

Mayr, E. (1942). *Systematics and the origins of species*. New York, NY: Columbia University Press.

Mendes, F. K., Hahn, Y., & Hahn, M. W. (2016). Gene tree discordance can generate patterns of diminishing convergence over time. *Molecular Biology and Evolution*, *33*(12), 3299–3307. https://doi.org/10.1093/molbev/msw197

Mergeay, J., & Santamaria, L. (2012). Evolution and Biodiversity: The evolutionary basis of biodiversity and its potential for adaptation to global change. *Evolutionary Applications*, *5*(2), 103–106. https://doi.org/10.1111/j.1752-4571.2011.00232.x

Merritt, T. J. S., & Bewick, A. J. (2017). Genetic diversity in insect metal tolerance. *Frontiers in Genetics*, *8*, 172. https://doi.org/10.3389/fgene.2017.00172

Migula, P., Wilczek, G., & Babczyńska, A. (2013). *Effects of heavy metal contamination. Spider ecophysiology*. Berlin, Heidelberg, Germany: Springer.

Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., & Kosakovsky Pond, S. L. (2012). Detecting individual sites subject to episodic diversifying selection. *PLoS Genetics*, *8*(7), e1002764. https://doi.org/10.1371/journal.pgen.1002764

Muschick, M., Indermaur, A., & Salzburger, W. (2012). Convergent evolution within an adaptive radiation of cichlid fishes. *Current Biology*, *22*(24), 2362–2368. https://doi.org/10.1016/J.CUB.2012.10.048

Nelson, D. R., & Nebert, D. W. (2011). *Cytochrome P450 (CYP) gene superfamily. Encyclopedia of life sciences*. Cichester, UK: John Wiley & Sons.

Norgate, M., Southon, A., Greenough, M., Cater, M., Farlow, A., Batterham, P., … Camakaris, J. (2010). Syntaxin 5 Is required for copper homeostasis in *Drosophila* and mammals. *PLoS ONE*, *5*(12), e14303. https://doi.org/10.1371/journal.pone.0014303

Nosil, P., Villoutreix, R., de Carvalho, C. F., Farkas, T. E., Soria-Carrasco, V., Feder, J. L., … Gompert, Z. (2018). Natural selection and

the predictability of evolution in Timema stick insects. *Science*, *359*(6377), 765–770. https://doi.org/10.1126/science.aap9125

Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, *302*(1), 205–217. https://doi.org/10.1006/jmbi.2000.4042

Paoletti, M. G., & Hassall, M. (1999). Woodlice (Isopoda: Oniscidea): Their potential for assessing sustainability and use as bioindicators. *Agriculture, Ecosystems & Environment*, *74*(1–3), 157–165. https://doi.org/10.1016/S0167-8809(99)00035-3

Pease, J. B., & Hahn, M. W. (2015). Detection and polarization of introgression in a five-taxon phylogeny. *Systematic Biology*, *64*(4), 651–662. https://doi.org/10.1093/sysbio/syv023

Pekár, S., Líznarová, E., Bočánek, O., & Zdráhal, Z. (2018). Venom of prey-specialized spiders is more toxic to their preferred prey: A result of prey-specific toxins. *Journal of Animal Ecology*, *87*(6), 1639–1652. https://doi.org/10.1111/1365-2656.12900

Pekár, S., Líznarová, E., & Řezáč, M. (2016). Suitability of woodlice prey for generalist and specialist spider predators: A comparative study. *Ecological Entomology*, *41*(2), 123–130. https://doi.org/10.1111/een.12285

Pond, S. L. K., Frost, S. D. W., & Muse, S. V. (2005). HyPhy: Hypothesis testing using phylogenies. *Bioinformatics*, *21*(5), 676–679. https://doi.org/10.1093/bioinformatics/bti079

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood trees for large alignments. *PLoS ONE*, *5*(3), e9490. https://doi.org/10.1371/journal.pone.0009490

Rey, C., Guéguen, L., Sémon, M., & Boussau, B. (2018). Accurate detection of convergent amino-acid evolution with PCOC. *Molecular Biology and Evolution*, *35*(9), 2296–2306. https://doi.org/10.1093/molbev/msy114

Řezáč, M., & Pekár, S. (2007). Evidence for woodlice-specialization in *Dysdera* spiders: Behavioural versus developmental approaches. *Physiological Entomology*, *32*(4), 367–371. https://doi.org/10.1111/j.1365-3032.2007.00588.x

Řezáč, M., Pekár, S., & Lubin, Y. (2008). How oniscophagous spiders overcome woodlouse armour. *Journal of Zoology*, *275*(1), 64–71. https://doi.org/10.1111/j.1469-7998.2007.00408.x

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). EDGER: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Roelofs, D., Janssens, T. K. S., Timmermans, M. J. T. N., Nota, B., MariËn, J., Bochdanovits, Z., … Van Straalen, N. M. (2009). Adaptive differences in gene expression associated with heavy metal tolerance in the soil arthropod *Orchesella cincta*. *Molecular Ecology*, *18*(15), 3227–3239. https://doi.org/10.1111/j.1365-294X.2009.04261.x

Rosenblum, E. B., Parent, C. E., & Brandt, E. E. (2014). The molecular basis of phenotypic convergence. *Annual Review of Ecology, Evolution, and Systematics*, *45*(1), 203–226. https://doi.org/10.1146/annurev-ecolsys-120213-091851

Sánchez-Herrero, J. F., Frías-López, C., Escuer, P., Hinojosa-Alvarez, S., Arnedo, M. A., Sánchez-Gracia, A., & Rozas, J. (2019). The draft genome sequence of the spider *Dysdera silvatica* (Araneae, Dysderidae): A valuable resource for functional and evolutionary genomic studies in chelicerates. *GigaScience*, In press.

Sanderson, M. J. (2002). Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. *Molecular Biology and Evolution*, *19*(1), 101–109. https://doi.org/10.1093/oxfordjournals.molbev.a003974

Sanderson, M. J. (2003). r8s: Inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, *19*(2), 301–302. https://doi.org/10.1093/bioinformatics/19.2.301

Schieber, M., & Chandel, N. S. (2014). ROS function in redox signaling and oxidative stress. *Current Biology*, *24*(10), R453–R462. https://doi.org/10.1016/j.cub.2014.03.034

Schluter, D. (2000). *The ecology of adaptive radiation*. Oxford, UK: Oxford University Press.

Schluter, D., & Conte, G. L. (2009). Genetics and ecological speciation. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(Supplement_1), 9955–9962. https://doi.org/10.1073/pnas.0901264106

Schmalfuss, H. (1984). Eco-morphological strategies in terrestrial isopods. *Symposium of the Zoological Society of London*, *53*, 49–63.

Smith, M. D., Wertheim, J. O., Weaver, S., Murrell, B., Scheffler, K., & Kosakovsky Pond, S. L. (2015). Less is more: An adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Molecular Biology and Evolution*, *32*(5), 1342–1353. https://doi.org/10.1093/molbev/msv022

Sooksa-Nguan, T., Yakubov, B., Kozlovskyy, V. I., Barkume, C. M., Howe, K. J., Thannhauser, T. W., … Vatamaniuk, O. K. (2009). *Drosophila* ABC transporter, DmHMT-1, confers tolerance to cadmium. DmHMT-1 and its yeast homolog, SpHMT-1, are not essential for vacuolar phytochelatin sequestration. *The Journal of Biological Chemistry*, *284*(1), 354–362. https://doi.org/10.1074/jbc.M806501200

Stamatakis, A. (2014). RAXML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Stern, D. L. (2013). The genetic causes of convergent evolution. *Nature Reviews Genetics*, *14*(11), 751–764. https://doi.org/10.1038/nrg3483

Storz, J. F. (2016). Causes of molecular convergence and parallelism in protein evolution. *Nature Reviews Genetics*, *17*(4), 239–250. https://doi.org/10.1038/nrg.2016.11

Stroud, J. T., & Losos, J. B. (2016). Ecological opportunity and adaptive radiation. *Annual Review of Ecology, Evolution, and Systematics*, *47*(1), 507–532. https://doi.org/10.1146/annurev-ecolsys-121415-032254

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE*, *6*(7), e21800. https://doi.org/10.1371/journal.pone.0021800

Sutton, S. L. (1980). *Woodlice*. New York, NY: Pergamon Press.

Toft, S., & Macías-Hernández, N. (2017). Metabolic adaptations for isopod specialization in three species of *Dysdera* spiders from the Canary Islands. *Physiological Entomology*, *42*(2), 191–198. https://doi.org/10.1111/phen.12192

Van Belleghem, S. M., Vangestel, C., De Wolf, K., De Corte, Z., Möst, M., Rastas, P., … Hendrickx, F. (2018). Evolution at two time frames: Polymorphisms from an ancient singular divergence event fuel contemporary parallel evolution. *PLOS Genetics*, *14*(11), e1007796. https://doi.org/10.1371/journal.pgen.1007796

van Ooik, T., & Rantala, M. J. (2010). Local adaptation of an insect herbivore to a heavy metal contaminated environment. *Annales Zoologici Fennici*, *47*(3), 215–222. https://doi.org/10.5735/086.047.0306

Van Straalen, N. M., & Roelofs, D. (2005). Cadmium tolerance in a soil arthropod a model of real-time microevolution. *Entomologische Berichten*, *65*(4), 105–111.

Vizueta, J., Frías-López, C., Macías-Hernández, N., Arnedo, M. A., Sánchez-Gracia, A., & Rozas, J. (2017). Evolution of chemosensory gene families in arthropods: Insight from the first inclusive comparative transcriptome analysis across spider appendages. *Genome Biology and Evolution*, *9*(1), 178–196. https://doi.org/10.1093/gbe/evw296

Vizueta, J., Rozas, J., & Sánchez-Gracia, A. (2018). Comparative genomics reveals thousands of novel chemosensory genes and massive changes in chemoreceptor repertories across chelicerates. *Genome Biology and Evolution*, *10*(5), 1221–1236. https://doi.org/10.1093/gbe/evy081

Wertheim, J. O., Murrell, B., Smith, M. D., Kosakovsky Pond, S. L., & Scheffler, K. (2015). RELAX: Detecting relaxed selection in a

phylogenetic framework. *Molecular Biology and Evolution*, *32*(3), 820–832. https://doi.org/10.1093/molbev/msu400

Whittaker, R. J., & Fernández-Palacios, J. M. (2007). *Island biogeography: Ecology, evolution, and conservation*. Oxford, UK: Oxford Univ. Press.

Wilczek, G., Babczyńska, A., Migula, P., & Wencelis, B. (2003). Activity of esterases as biomarkers of metal exposure in spiders from the metal pollution gradient. *Polish Journal of Environmental Studies*, *12*(6), 765–771.

World Spider Catalog (2019). *World spider catalog. Version 20.0*. Bern, Switzerland: Natural History Museum Bern. Retrieved from http://wsc.nmbe.ch

Wu, M., Kostyun, J. L., Hahn, M. W., & Moyle, L. C. (2018). Dissecting the basis of novel trait evolution in a radiation with widespread phylogenetic discordance. *Molecular Ecology*, *27*(16), 3301–3316. https://doi.org/10.1111/mec.14780

Zapata, M., Tanguy, A., David, E., Moraga, D., & Riquelme, C. (2009). Transcriptomic response of *Argopecten purpuratus* post-larvae to copper exposure under experimental conditions. *Gene*, *442*(1–2), 37–46. https://doi.org/10.1016/J.GENE.2009.04.019

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, *19*(S6), 153. https://doi.org/10.1186/s12859-018-2129-y

Zhang, Y., Lambiase, S., Fasola, M., Gandini, C., Grigolo, A., & Laudani, U. (2001). Mortality and tissue damage by heavy metal contamination in the German cockroach, *Blattella germanica* (Blattaria, Blattellidae). *Italian Journal of Zoology*, *68*(2), 137–145. https://doi.org/10.1080/11250000109356398

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Vizueta J, Macías-Hernández N, Arnedo MA, Rozas J, Sánchez-Gracia A. Chance and predictability in evolution: The genomic basis of convergent dietary specializations in an adaptive radiation. *Mol Ecol*. 2019;28:4028–4045. https://doi.org/10.1111/mec.15199

Chance and predictability in evolution: the genomic basis of convergent dietary specializations in an adaptive radiation

Vizueta J., Macías-Hernández N., Arnedo M.A., Rozas J. and Sánchez-Gracia A.

Supplementary Material

MOLECULAR ECOLOGY

Supplemental Information for:

Chance and predictability in evolution: the genomic basis of convergent dietary specializations in an adaptive radiation

Joel Vizueta[1], Nuria Macías-Hernández[2, 3], Miquel A. Arnedo[4], Julio Rozas[1]* and Alejandro Sánchez-Gracia[1]*

[1]Departament de Genètica, Microbiologia i Estadística, and Institut de Recerca de la Biodiversitat (IRBio), Facultat de Biologia, Universitat de Barcelona, Diagonal 643, 08028, Barcelona, Spain.

[2]Laboratory for Integrative Biodiversity Research, Finnish Museum of Natural History, University of Helsinki; PO Box 17, 00014 Helsinki, Finland.

[3]Island Ecology and Evolution Research Group, Instituto de Productos Naturales y Agrobiología (IPNA-CSIC). C/Astrofísico Francisco Sánchez 3. La Laguna, Tenerife, Canary Islands, 38206, Spain

[4]Departament de Biologia Evolutiva, Ecologia i Ciències Ambientals and Institut de Recerca de la Biodiversitat (IRBio), Facultat de Biologia, Universitat de Barcelona, Diagonal 643, 08028, Barcelona, Spain

* Corresponding authors. E-mail: jrozas@ub.edu, elsanchez@ub.edu

MOLECULAR ECOLOGY

Supplementary Methods, Tables & Figures

Supplementary Methods

Transcriptome assembly and functional annotation

We used NGSQCToolkit (Patel & Jain, 2012) to remove low quality reads (reads with more than 30% of bases with quality scores < 20) and reads containing adaptors and missing data. Filtered reads were further corrected for sequencing errors with SEECER v_0.1.3 (Le, Schulz, McCauley, Hinman, & Bar-Joseph, 2013) and *de novo* assembled using Bridger with *k*-mer size of 31 (Chang et al., 2015). After discarding contaminant contigs with Seqclean software (http://compbio.dfci.harvard.edu/tgi/software/), we clustered the contigs into individual transcripts or components. We assessed the transcriptome completeness in each species by estimating the percentage of Core Eukaryotic Genes (CEG) encoding transcripts present in the four assemblies (TBLASTN searches against the CEG database (Parra, Bradnam, & Korf, 2007); *E*-value < 10^{-5}).

For the functional annotation of new generated transcripts we carried out exhaustive BLAST searches against NCBI-nr and ArthropodDB databases (including well annotated genomes of related chelicerate species (e.g. Gulia-Nuss et al., 2016; Hoy et al., 2016; Schwager et al., 2017; see Vizueta et al., 2017 for further details). We also used Pfam profiles (Finn et al., 2014) (curated hidden markov models, HMM) as queries in HMMER (Eddy, 2011) searches against translated proteins and InterProScan (Jones et al., 2014) to detect protein-domain signatures in translated sequences. Gene ontology (GO) terms (Ashburner et al., 2000) for each transcript were inherited from the best significant blast or HMMER results (E-value < 10^{-5}). Coding sequences (CDS) were inferred by combining the results of TransDecoder tool (Haas et al., 2013), which predicts opening reading frames (ORF) in the transcripts, and of the BLAST searches against the above mentioned databases. Lastly, we collapsed translated protein sequences showing high sequence similarity using CD-HIT (Fu, Niu, Zhu, Wu, & Li, 2012) (99% identity), to remove putative isoforms.

We also specifically searched the five transcriptomes for sequences encoding spider venom components, which would be important candidates to be involved in dietary prey specialization. We used all translated peptides as a query in a BLASTP search against the Araneomorphae sequences in ArachnoServer (Pineda et al., 2018).

Differential expression analyses in chemosensory gene families

Given the difficulty to conduct a fine determination of the number of copies (and the orthogroups) of gene families from a transcriptome obtained after the *de novo* assembly of short reads, we did not analyze their DE patterns, with the exception of chemosensory gene family members expressed in spider chemosensory appendages. Given that chemosensory system is likely involved in prey detection and avoidance (Sánchez-Gracia, Vieira, & Rozas, 2009), the members of these families (GR, IR, CD36-SNMP, OBP-like, NPC2 and CCP families; Vizueta et al., 2018) are, a priori, firm candidates to have undergone convergent adaptations during dietary specializations of Canarian *Dysdera*. In order to approximate the DE profiles in chemosensory genes, we first annotated the transcripts and estimated their phylogenetic relationships within each family to identify possible single-copy orthologs.

Multiple sequence alignments for selective constraints analyses

We aligned the CDS of all single-copy orthologs of the using the software PRANK (Loytynoja & Goldman, 2008). This software has been shown to perform accurate MSA of coding sequences, especially suitable for the analysis of selective constraints (Jordan & Goldman, 2012). Only the orthologous groups with no evidence of gene conversion and with good ZORRO confidence scores were included in this analysis (Sawyer, 1989; Wu, Chatterji, Eisen, Glaser, & Ben-Tal, 2012) (Figure 2).

MOLECULAR ECOLOGY

References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, *25*(1), 25–29. doi:10.1038/75556

Chang, Z., Li, G., Liu, J., Zhang, Y., Ashby, C., Liu, D., … Huang, X. (2015). Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. *Genome Biology*, *16*(1), 1–10. doi:10.1186/s13059-015-0596-2

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, *7*(10), e1002195. doi:10.1371/journal.pcbi.1002195

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., … Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Research*, *42*, D222–D230. doi:10.1093/nar/gkt1223

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152. doi:10.1093/bioinformatics/bts565

Gulia-Nuss, M., Nuss, A. B., Meyer, J. M., Sonenshine, D. E., Roe, R. M., Waterhouse, R. M., … Hill, C. A. (2016). Genomic insights into the *Ixodes scapularis* tick vector of Lyme disease. *Nature Communications*, *7*, 10507. doi:10.1038/ncomms10507

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512. doi:10.1038/nprot.2013.084

Hoy, M. A., Waterhouse, R. M., Wu, K., Estep, A. S., Ioannidis, P., Palmer, W. J., … Richards, S. (2016). Genome Sequencing of the Phytoseiid Predatory Mite *Metaseiulus occidentalis*

Reveals Completely Atomized Hox Genes and Superdynamic Intron Evolution. *Genome Biology and Evolution*, *8*(6), 1762–1775. doi:10.1093/gbe/evw048

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., … Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240. doi:10.1093/bioinformatics/btu031

Jordan, G., & Goldman, N. (2012). The Effects of Alignment Error and Alignment Filtering on the Sitewise Detection of Positive Selection. *Molecular Biology and Evolution*, *29*(4), 1125–1139. doi:10.1093/molbev/msr272

Le, H.-S., Schulz, M. H., McCauley, B. M., Hinman, V. F., & Bar-Joseph, Z. (2013). Probabilistic error correction for RNA sequencing. *Nucleic Acids Research*, *41*(10), e109. doi:10.1093/nar/gkt215

Loytynoja, A., & Goldman, N. (2008). Phylogeny-Aware Gap Placement Prevents Errors in Sequence Alignment and Evolutionary Analysis. *Science*, *320*(5883), 1632–1635. doi:10.1126/science.1158395

Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, *23*(9), 1061–1067. doi:10.1093/bioinformatics/btm071

Patel, R. K., & Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PloS One*, *7*(2), e30619. doi:10.1371/journal.pone.0030619

Pineda, S. S., Chaumeil, P.-A., Kunert, A., Kaas, Q., Thang, M. W. C., Le, L., … King, G. F. (2018). ArachnoServer 3.0: an online resource for automated discovery, analysis and annotation of spider toxins. *Bioinformatics*, *34*(6), 1074–1076. doi:10.1093/bioinformatics/btx661

Sánchez-Gracia, A., Vieira, F. G., & Rozas, J. (2009). Molecular evolution of the major

chemosensory gene families in insects. *Heredity*, *103*(3), 208–216. doi:10.1038/hdy.2009.55

Sawyer, S. (1989). Statistical tests for detecting gene conversion. *Molecular Biology and Evolution*, *6*(5), 526–538.

Schwager, E. E., Sharma, P. P., Clarke, T., Leite, D. J., Wierschin, T., Pechmann, M., … McGregor, A. P. (2017). The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. *BMC Biology*, *15*(1), 62. doi:10.1186/s12915-017-0399-x

Vizueta, J., Frías-López, C., Macías-Hernández, N., Arnedo, M. A., Sánchez-Gracia, A., & Rozas, J. (2017). Evolution of chemosensory gene families in arthropods: Insight from the first inclusive comparative transcriptome analysis across spider appendages. *Genome Biology and Evolution*, *9*(1), 178–196. doi:10.1093/gbe/evw296

Vizueta, J., Rozas, J., & Sánchez-Gracia, A. (2018). Comparative Genomics Reveals Thousands of Novel Chemosensory Genes and Massive Changes in Chemoreceptor Repertories across Chelicerates. *Genome Biology and Evolution*, *10*(5), 1221–1236. doi:10.1093/gbe/evy081

Wu, M., Chatterji, S., Eisen, J. A., Glaser, F., & Ben-Tal, N. (2012). Accounting For Alignment Uncertainty in Phylogenomics. *PLoS ONE*, *7*(1), e30288. doi:10.1371/journal.pone.0030288

MOLECULAR ECOLOGY

Supplementary figures

Figure S1. Distribution of blastx hits across species. Distribution of the top 5 hits from the blastx searches with the transcripts of each *Dysdera* species against the ArthropodDB database.

Figure S2. Principal component analysis (PCA) of gene expression profiles of individual *REST* samples from *D. tilosensis*.

Figure S3. Venn diagrams showing (**a**) the number of shared genes between species pairs. Differential expressed (DE) genes are showed in brackets; (**b**) the number of DE genes between species pairs and groups of tissues (*LEGS-PALP* refers to the *LEG#1*, *LEG#234* and *PALP*); (**c**) number of MGE candidates across tissues.

Figure S4. Tree maps with detailed GO enrichment results generated with REVIGO.

Figure S5. Box plots showing the distribution of ω values for all single-copy orthogroups in specialist (orange) and generalist (blue) species.

Figure S6. Orthogroups with evidence of convergent amino acid evolution. Amino acid positions are coloured according to their profiles, and only positions with a *PP* equal to or greater than 0.99 according to the PCOC, PC or OC model are shown. Yellow stars highlight the sites identified as positively selected in MEME.

MOLECULAR ECOLOGY

Supplementary tables

Table S1. RNA-seq statistics.

Table S2. Distribution of the percentage of CEG length covered by blastx hits.

Table S3. Orthogroups classification.

Table S4. List of genes with concordant differential expression profiles between generalist and specialist species.

Table S5. List of genes with concordant differential functional constraint profiles between generalist and specialist species.

Table S6. List of genes with concordant signals of positive selection in specialist species.

Table S1. RNA-seq statistics

Species	Sample	Number of individuals	Sex	Capture Date	Location	Geographic coordinates	Habitat	Spiders ID	Total bases	Read count	GC(%)	Q20(%)
Dysdera silvatica	*LEG#1*	*4*	*Male*	May 2013	*Las Tajoras, La Gomera*	*28.11273 -17.26251*	*Laurel forest*	*NMH2597, NMH2598, NMH2599, NMH2601*	11,919,755,986	118,017,386	41.38	96.78
Dysdera silvatica	*PALP*	*4*	*Male*	May 2013	*Las Tajoras, La Gomera*	*28.11273 -17.26251*	*Laurel forest*	*NMH2597, NMH2598, NMH2599, NMH2601*	11,613,604,382	114,986,182	41.41	96.39
Dysdera silvatica	*LEG#234*	*4*	*Male*	May 2013	*Las Tajoras, La Gomera*	*28.11273 -17.26251*	*Laurel forest*	*NMH2597, NMH2598, NMH2599, NMH2601*	10,490,369,040	103,865,040	41.55	96.63
Dysdera silvatica	*REST*	*4*	*Male*	May 2013	*Las Tajoras, La Gomera*	*28.11273 -17.26251*	*Laurel forest*	*NMH2597, NMH2598, NMH2599, NMH2601*	10,601,692,856	104,967,256	41.39	96.53
Dysdera verneaui	*LEG#1*	*12*	*Male*	May 2013	*Pista Batán. Cruz del Carmen. Tenerife*	*28.53532 -16.2968*	*Laurel forest*	*12580, NMH2582, NMH2583, NMH2585, NMH2586, NMH2*	11,548,986,804	114,346,404	39.33	95.77
Dysdera verneaui	*PALP*	*3*	*Male*	May 2013	*Pista Batán. Cruz del Carmen. Tenerife*	*28.53532 -16.2969*	*Laurel forest*	*NMH2585, NMH2586, NMH2587*	11,050,404,344	109,409,944	37.34	95.17
Dysdera verneaui	*LEG#234*	*4*	*Male*	May 2013	*Pista Batán. Cruz del Carmen. Tenerife*	*28.53532 -16.2968*	*Laurel forest*	*NMH2580, NMH2581, NMH2582, NMH2583*	12,626,674,580	125,016,580	41.00	96.28
Dysdera verneaui	*REST*	*3*	*Male*	May 2013	*Pista Batán. Cruz del Carmen. Tenerife*	*28.53532 -16.2969*	*Laurel forest*	*NMH2570, NMH2579, NMH2581*	10,008,708,524	99,096,124	38.00	94.08
Dysdera verneaui	*REST 2*	*2*	*Male*	May 2013	*Pista Batán. Cruz del Carmen. Tenerife*	*28.53532 -16.2968*	*Laurel forest*	*NMH2586, NMH2587*	18,031,153,674	178,526,274	39.02	97.89
Dysdera gomerensis	*LEG#1*	*8*	*Female*	May 2013	*Ventejís, El Hierro*	*27.78599 -17.93750*	*Pine forest*	*12540, NMH2541, NMH2542, NMH2543, NMH2544, NMH2*	9,965,727,974	98,670,574	39.59	95.75
Dysdera gomerensis	*PALP*	*8*	*Female*	May 2013	*Ventejís, El Hierro*	*27.78599 -17.93750*	*Pine forest*	*12540, NMH2541, NMH2542, NMH2543, NMH2544, NMH2*	9,780,095,428	96,832,628	40.37	95.43
Dysdera gomerensis	*LEG#234*	*7*	*Female*	May 2013	*Ventejís, El Hierro*	*27.78599 -17.93750*	*Pine forest*	*, NMH2547, NMH2558, NMH2557, NMH2558, NMH2559,*	11,751,524,326	116,351,726	40.78	95.63
Dysdera gomerensis	*REST*	*3*	*Female*	May 2013	*Ventejís, El Hierro*	*27.78599 -17.93750*	*Pine forest*	*NMH2540, NMH2541, NMH2542*	11,985,398,512	118,667,312	40.92	96.39
Dysdera bandamae	*LEG#1*	*13*	*Male and Female*	May 2015	*Llanos de la Pez, Gran Canaria*	*27.96431 -15.58554*	*Pine forest*	*CRBA2177, CRBA2178, CRBA2179, CRBA2180, CRBA2176*	17,176,194,330	170,061,330	35.06	97.748
Dysdera bandamae	*PALP*	*14*	*Male and Female*	May 2015	*Llanos de la Pez, Gran Canaria*	*27.96431 -15.58554*	*Pine forest*	*2189, CRBA2177, CRBA2178, CRBA2179, CRBA2180, CRB*	15,486,553,816	153,332,216	35.51	97.654
Dysdera bandamae	*LEG#234*	*14*	*Male and Female*	May 2015	*Llanos de la Pez, Gran Canaria*	*27.96431 -15.58554*	*Pine forest*	*2189, CRBA2177, CRBA2178, CRBA2179, CRBA2180, CRB*	14,896,839,864	147,493,464	34.91	97.544
Dysdera bandamae	*REST M1*	*1*	*Male*	May 2015	*Llanos de la Pez, Gran Canaria*	*27.96431 -15.58554*	*Pine forest*	*CRBA2181*	8,102,388,468	80,221,668	35.65	97.738
Dysdera bandamae	*REST M2*	*1*	*Male*	May 2015	*Llanos de la Pez, Gran Canaria*	*27.96431 -15.58554*	*Pine forest*	*CRBA2182*	7,063,940,202	69,940,002	34.81	97.012
Dysdera bandamae	*REST F1*	*1*	*Female*	May 2015	*Llanos de la Pez, Gran Canaria*	*27.96431 -15.58554*	*Pine forest*	*CRBA2185*	7,455,863,228	73,820,428	35.44	97.467
Dysdera bandamae	*REST F2*	*1*	*Female*	May 2015	*Llanos de la Pez, Gran Canaria*	*27.96431 -15.58554*	*Pine forest*	*CRBA2178*	7,149,219,754	70,784,354	36.73	97.934
Dysdera tilosensis	*LEG#1*	*5*	*Male*	May 2015	*Degollada de Becerra, Presa de Cuevas Blancas, Cor*	*27.98973 -15.59292; 2*	*Pine forest*	*CRBA2219, CRBA2220, NMH3147, NMH3154, NMH3155*	6,571,489,250	65,064,250	39.81	97.45
Dysdera tilosensis	*PALP*	*5*	*Male*	May 2015	*Degollada de Becerra, Presa de Cuevas Blancas, Cor*	*27.98973 -15.59292; 2*	*Pine forest*	*CRBA2219, CRBA2220, NMH3147, NMH3154, NMH3155*	6,919,567,368	68,510,568	39.42	97.85
Dysdera tilosensis	*LEG#234*	*15*	*Male and Female*	May 2015	*Degollada de Becerra, Presa de Cuevas Blancas, Cor*	*27.98973 -15.59292; 2*	*Pine forest*	*CRBA2195, CRBA2597, CRBA2598, CRBA2201, CRBA2194*	5,919,320,130	58,607,130	42.00	98.29
Dysdera tilosensis	*REST M1*	*1*	*Male*	May 2015	*Degollada de Becerra, Gran Canaria*	*27.98973 -15.59293*	*Pine forest*	*CRBA2195*	5,193,796,326	51,423,726	39.44	97.98
Dysdera tilosensis	*REST M2*	*1*	*Male*	May 2015	*Degollada de Becerra, Gran Canaria*	*27.98973 -15.59292*	*Pine forest*	*CRBA2598*	6,483,209,190	64,190,190	40.96	98.1
Dysdera tilosensis	*REST M3*	*1*	*Male*	May 2015	*Presa de Cuevas Blancas, Gran Canaria*	*27.96483 -15.54505*	*Pine forest*	*CRBA2200*	6,512,922,178	64,484,378	39.39	98.05
Dysdera tilosensis	*REST M4*	*1*	*Male*	May 2015	*Cortijo San Gregorio, Tamaraceite, Gran Canaria*	*28.08578 -15.49128*	*Pine forest*	*NMH3147*	6,659,775,370	65,938,370	41.05	98.19
Dysdera tilosensis	*REST F1*	*1*	*Female*	May 2015	*Degollada de Becerra, Gran Canaria*	*27.98973 -15.59293*	*Pine forest*	*CRBA2194*	4,400,839,670	43,572,670	41.24	98
Dysdera tilosensis	*REST F2*	*1*	*Female*	May 2015	*Degollada de Becerra, Gran Canaria*	*27.98973 -15.59292*	*Pine forest*	*CRBA2196*	6,221,035,208	61,594,408	41.63	98.16
Dysdera tilosensis	*REST F3*	*1*	*Female*	May 2015	*Presa de Cuevas Blancas, Gran Canaria*	*27.96483 -15.54505*	*Pine forest*	*CRBA2199*	7,201,008,918	71,297,118	40.11	97.71
Dysdera tilosensis	*REST F4*	*1*	*Female*	May 2015	*Cortijo San Gregorio, Tamaraceite, Gran Canaria*	*28.08578 -15.49128*	*Pine forest*	*MNH3146*	6,410,225,176	63,467,576	40.61	98.09

Each row corresponds to an independent RNA sequencing experiment. Column C indicates the number of individuals sequenced as a whole

Table S2. Distribution of the percentage of CEG length covered by blastx hits.

Percentage of coverage	*D. silvatica*	*D. verneaui*	*D. gomerensis*	*D. bandamae*	*D. tilosensis*
Total	457	457	457	457	457
100	234	243	240	246	249
80	429	431	434	431	438
50	455	456	455	454	455

The complete CEG database includes 458 genes

Table S3. Orthogroups classification

Orthologues	GV pair	TB pair	*5 Dysdera*	*5 Dysdera + D. crocata*
	D. gomerensis - D. verneaui	*D. tilosensis - D. bandamae*		
Orthogroups	58,600	58,600	58,600	62,318
N:N orthogroups (N>= 1)	23,602	28,766	13,947	10,503
1:1 orthologs	19,497[a]	24,212[a]	9,473	6,575
1:1 without stop codons	-	-	7,958[b]	5,958
1:1 complete sequences	-	-	4,539[c]	2,472[d]
Gene families (excluding 1:1)	4,105	4,554	4,474[b]	3,928
Gene families without stop codons	-	-	2,442	-

[a] Used in differential expression analysis

[b] Used in functional constraint analysis

[c] 6,037,300 nucleotides. Used for divergence time estimation

[d] 2,926,723 nucleotides. Used in the phylogenomic analysis

Table S4. List of genes with concordant differential expression profiles between generalist and specialists species. Available at *Molecular Ecology* online https://doi.org/10.1111/mec.15199

Table S5. List of genes with concordant differential functional constraint profiles between generalist and specialist species.

Orthogroup	Selection in the specialist	Gene annotation	GO annotation
OG12289	Intensified	neuropeptide Y receptor-like	GO:0016021;GO:0004983;GO:0004871;GO:0004872;
OG14127	Intensified	Zinc finger protein	GO:0046872;GO:0003676;GO:0005622;GO:0008270;
OG7315	Intensified	membrane glycoprotein LIG-1	GO:0005515;
OG11425	Relaxed	hexokinase	GO:0005524;GO:0005975;GO:0016773;GO:0006096;
OG12249	Relaxed	zinc finger protein 62 homolog	GO:0046872;GO:0003676;GO:0005622;GO:0008270;
OG12775	Relaxed	Paired AMPhipathic helix protein Sin3A	GO:0005634;GO:0006355;
OG15821	Relaxed	conserved unknown protein	GO:0006913;GO:0008565;GO:0005488;
OG7594	Relaxed	KRAB domain-containing zinc finger protein	GO:0046872;GO:0003676;GO:0003677;GO:0005622;
OG8895	Relaxed	similar to interphase cyctoplasmic foci protein 45	GO:0000287;GO:0006400;GO:0008193;

Table S6. List of genes with concordant signals of positive selection in specialist species.

Orthogroup	Hemiplasy	Gene annotation	GO annotation	Predicted toxin	Convergent sites (PCOC)	Convergent sites (PCOC) under positive selection
OG6752		nidogen (entactin) cysteine-type endopeptidase inhibitor activity	GO:0004869;GO:0005576;GO:0010466;G	U24-ctenitoxin-Pn1a	6	0
OG10211		NA	NA		2	0
OG10499	Probable	troponin 1 [Latrodectus hesperus],gi|318087220|gb|ADV40202.1|,5	GO:0005861;		-	-
OG10868		CUFF.81277.2_Ste Uncharacterized protein (Fragment)L37654_T1/1_	NA		1	0
OG11238		class B secretin-like G-protein coupled receptor GPRmth5, putative |	GO:0005044;GO:0006955;GO:0030247;		0	0
OG11255		PREDICTED: mannose receptor C type 1-like [Saccoglossus kowalevsl	GO:0030246;GO:0005488;GO:0005529;		5	2
OG13286		PREDICTED: sodium channel, nonvoltage-gated 1 gamma-like [Sacco	GO:0005272;GO:0006814;GO:0016020;G		1	1
OG13706		translation initiation factor eIF-2B subunit delta [Oryctolagus cunicu	GO:0044237;GO:0003743;		2	0
OG14147	Probable	PREDICTED: tudor domain-containing protein 5-like [Anolis caroliner	GO:0016301;GO:0003723;		-	-
OG14203	Probable	membrane glycoprotein LIG-1, putative [Ixodes scapularis],gi|24117	GO:0005515;		-	-
OG16682		carbon-nitrogen hydrolase, putative [Ixodes scapularis],gi|24124562	GO:0006807;GO:0016810;GO:0016811;		4	1
OG16879	Probable	PREDICTED: methyltransferase-like protein 7B-like [Monodelphis do	GO:0008152;GO:0008168;		-	-
OG17087		NA	NA		1	0
OG6704		KOG1656-like protein [Ornithodoros moubata],gi|45386075|gb|AA5	GO:0015031;		1	0
OG6889		wunen-trimeric MYC tag fusion protein [synthetic construct],gi|275	GO:0003824;GO:0016020;GO:0008354;G		1	0
OG7181		PREDICTED: tectonin beta-propeller repeat-containing protein 1, pa	GO:0016021;		10	3
OG7495		PREDICTED: similar to CRAL/TRIO domain-containing protein [Triboli	GO:0005215;GO:0005622;GO:0006810;		0	0
OG8162		leucine rich repeat transmembrane neuronal 4 [Lycosa singoriensis],	GO:0005515;		2	0
OG8461		Ca2+ calmodulin dependent protein kinase EF-Hand protein superfa	GO:0005509;		0	0
OG9489	Probable	CRE-LEA-1 protein [Caenorhabditis remanei],gi|308508313|ref|XP_	GO:0005488;GO:0003677;GO:0005576;G		-	-
OG9529		PREDICTED: epidermal retinal dehydrogenase 2 [Taeniopygia guttata	GO:0016491;GO:0008152;		4	0
OG9641		LOC407707 protein, partial [Danio rerio],gi|60551198|gb|AAH9091	GO:0003677;GO:0020037;GO:0004601;G		3	1

Table S6. List of genes with concordant signals of positive selection in specialist species. Full tables available at *Molecular Ecology* online https://doi.org/10.1111/mec.15199

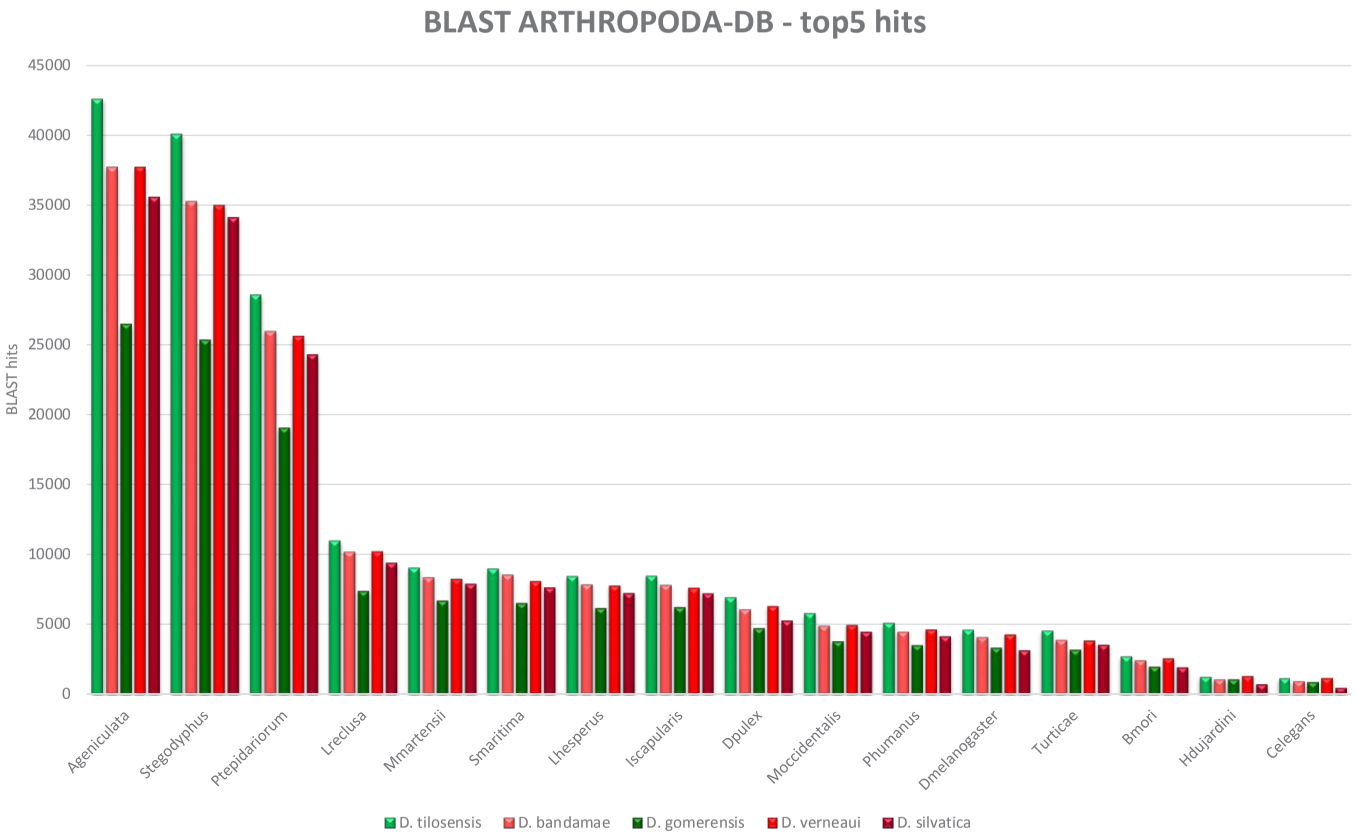


Figure S1. Distribution of blastx hits across species. Distribution of the top 5 hits from the blastx searches with the transcripts of each *Dysdera* species against the ArthropodDB database.

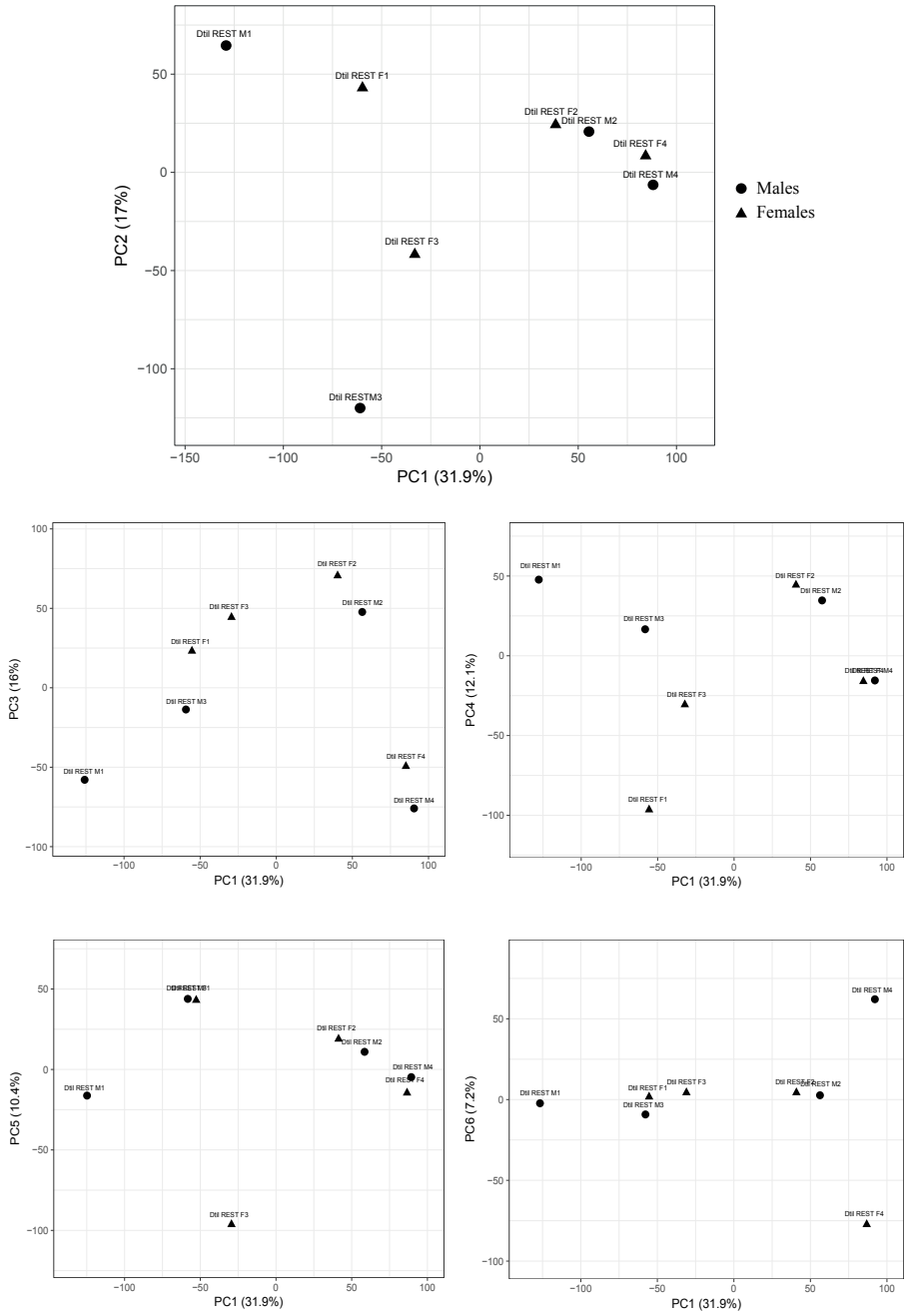
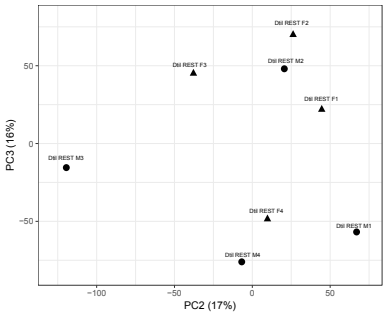
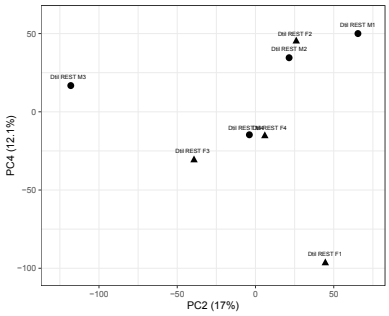


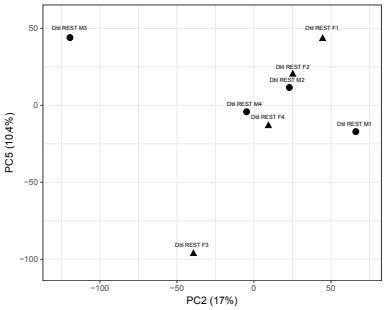
Figure S2. Principal component analysis (PCA) of gene expression profiles of individual *REST* samples from *D. tilosensis*.

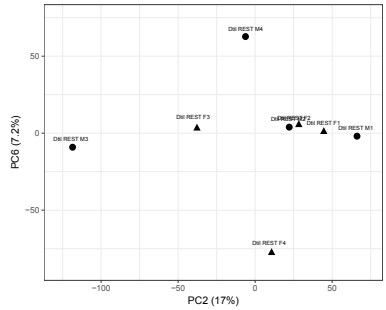
continued on next page

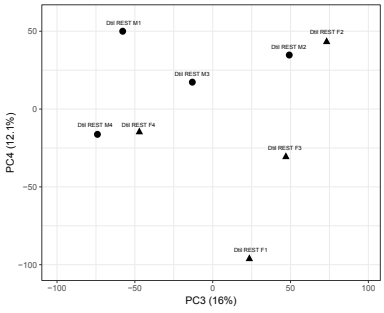
continued from previous page

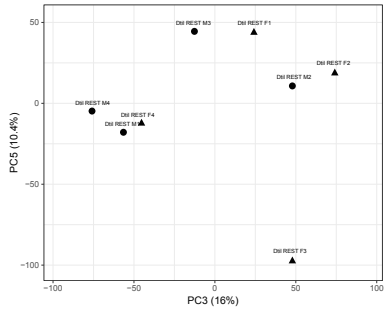


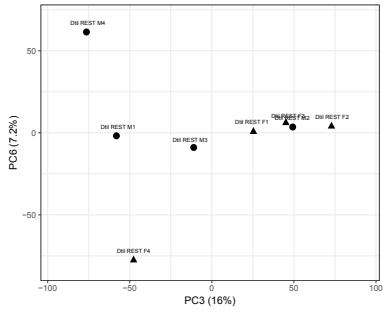


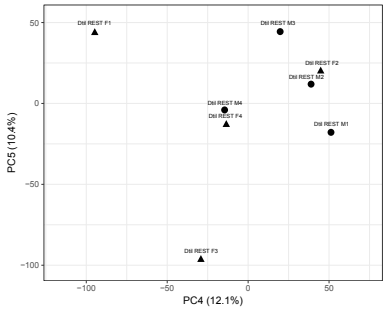












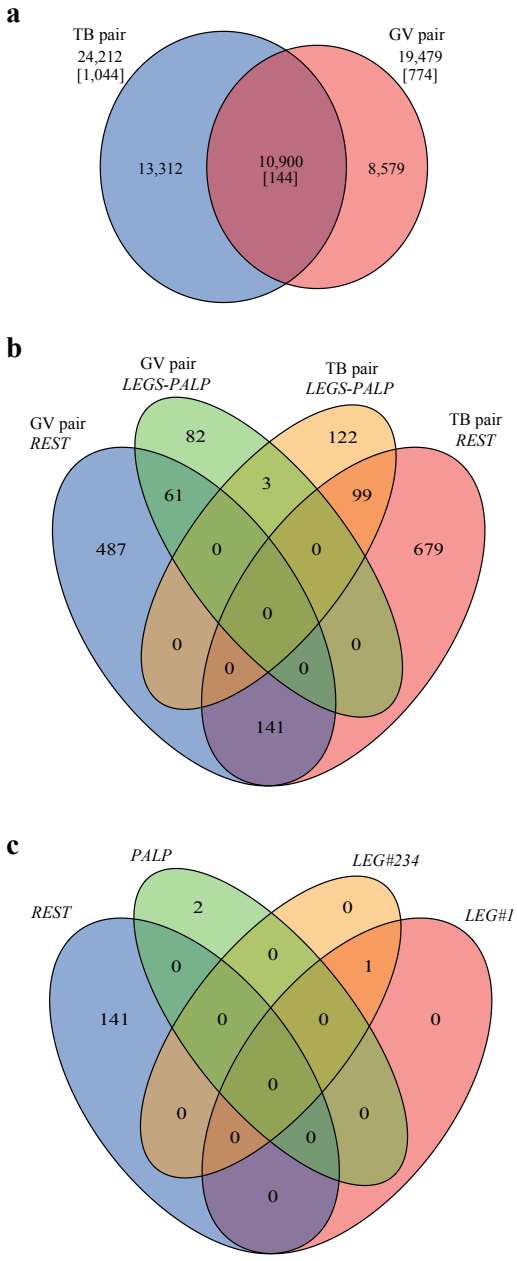
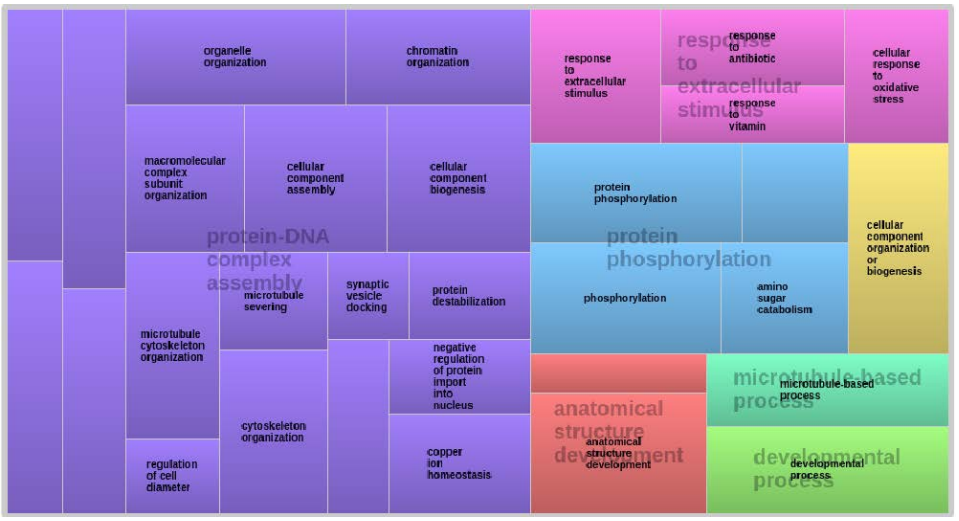


Figure S3. Venn diagrams showing (a) the number of shared genes between species pairs. Differential expressed (DE) genes are showed in brackets; (b) the number of DE genes between species pairs and groups of tissues (*LEGS-PALP* refers to the *LEG#1*, *LEG#234* and *PALP*); (c) number of MGE candidates across tissues.

DE Analysis

Common in both species pairs (GO enrichment)

Biological Process



Molecular Function

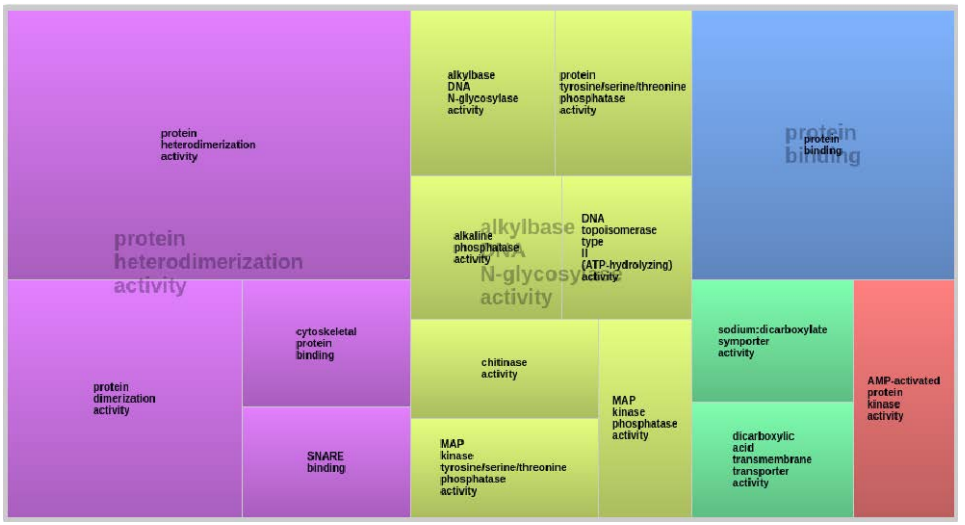


Figure S4. Tree maps with detailed GO enrichment results generated with REVIGO. Full figure available at *Molecular Ecology* online https://doi.org/10.1111/mec.15199

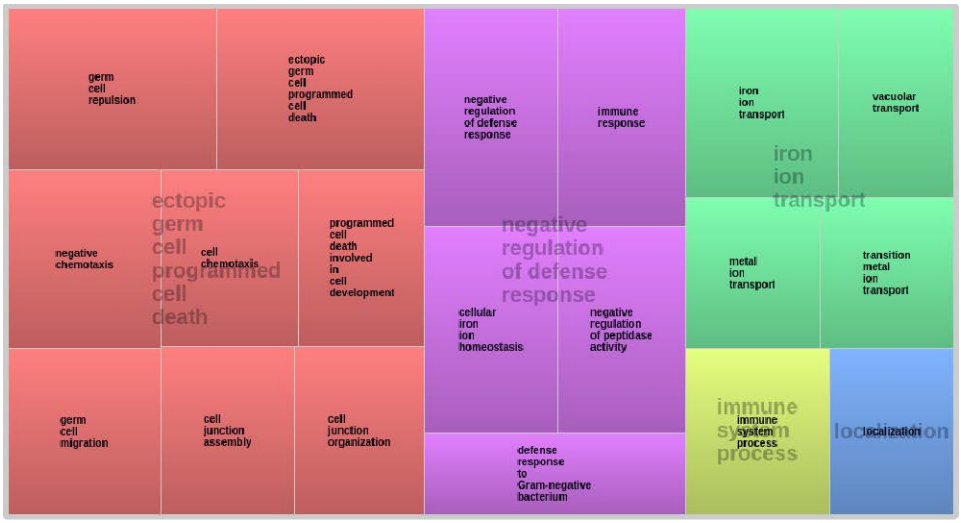
continued on next page

continued from previous page

PS Analysis

Common in both species pairs (GO enrichment)

Biological Process



Molecular Function

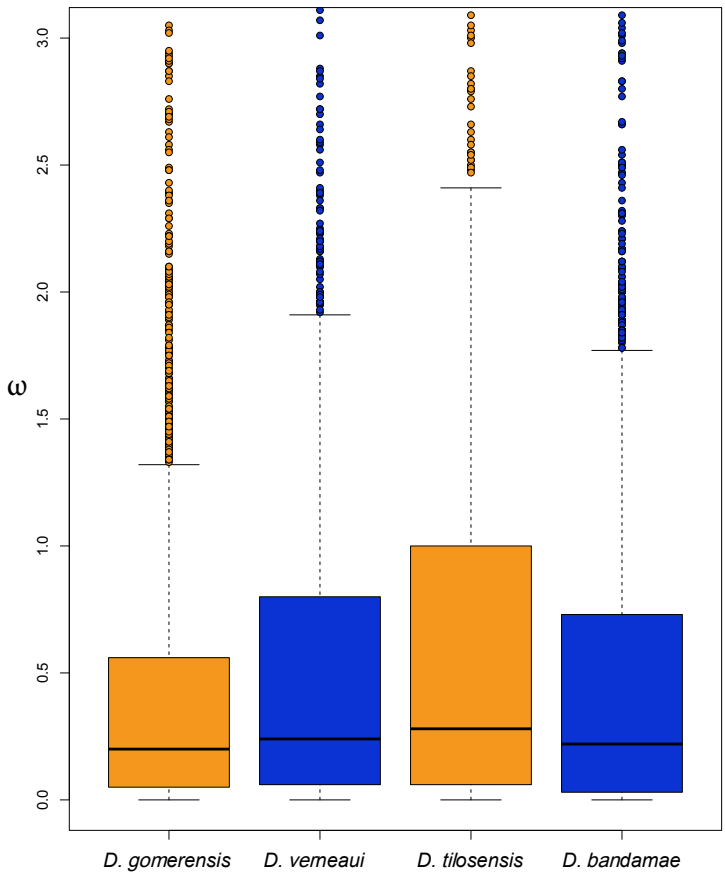


Figure S5. Box plots showing the distribution of ω values for all single-copy orthogroups in specialist (orange) and generalist (blue) species.

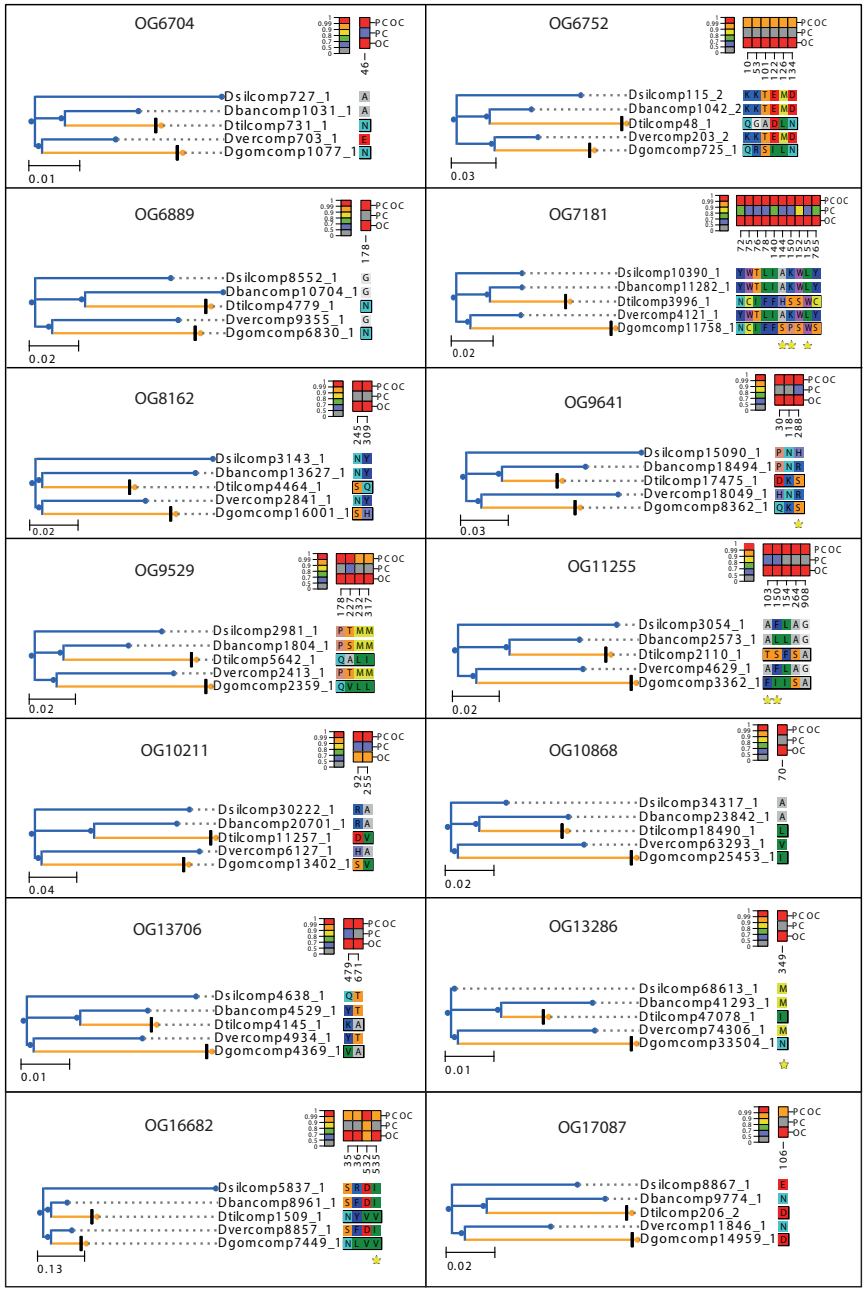


Figure S6. Orthogroups with evidence of convergent amino acid evolution. Amino acid positions are coloured according to their profiles, and only positions with a PP equal to or greater than 0.99 according to the PCOC, PC or OC model are shown. Yellow stars highlight the sites identified as positively selected in MEME.

Discusión

El desarrollo de las nuevas tecnologías de secuenciación (conocidas como *Next Generation Sequencing*, NGS) ha permitido pasar de analizar un locus o varios loci[143,144] a ser capaces de analizar todos los loci del genoma, tanto en organismos modelo como no modelo[145–148]. Este hecho ha significado un punto de inflexión en la genética evolutiva y ha promovido su transformación en genómica evolutiva, facilitando que estudios que no se podrían llevar a cabo con un solo locus, como por ejemplo discernir el papel relativo de la selección natural y la demografía en la generación de los patrones de variabilidad observados en las poblaciones naturales[149], ahora se puedan abordar gracias a la información genómica. Aunque ambos procesos pueden dejar huellas similares en el genoma, los efectos de la selección son mucho más locales (afectando a regiones concretas del genoma) mientras que las señales que deja la historia demográfica son mucho más amplias y afectan a todo el genoma (o a una región importante del mismo).

Por tanto, la disponibilidad de datos genómicos nos permite estudiar los efectos de la selección natural en la variación genética a lo largo del genoma, y de este modo intentar abordar distintas cuestiones de la biología y genética evolutiva[150]. Entre estas preguntas fundamentales encontramos la de identificar y determinar el efecto fenotípico de los genes y los tipos de mutaciones que son diana de la selección en los procesos adaptativos[149]. A priori, existen ciertos genes que son candidatos a evolucionar más frecuentemente bajo selección positiva que otros, como por ejemplo los genes relacionados con el sistema inmune o la respuesta a patógenos[151–153]. No obstante, la aproximación genómica permite delimitar y establecer la función de los genes implicados en respuestas adaptativas sin ninguna premisa previa, pudiendo así identificar nuevos determinantes genéticos (genes que son o han sido diana de la selección pero que no habían sido considerados como candidatos a priori)[154,155]. A su vez, estos análisis nos permiten conocer los tipos de mutaciones afectados por la selección natural (p. ej. mutaciones puntuales, inserciones o deleciones, inversiones, duplicaciones, etc..) además de su localización genómica y su potencial efecto fenotípico[149]. En esta tesis doctoral se han utilizado datos de secuenciación

masiva (RNA-seq y secuencias genómicas) para profundizar en el conocimiento de los procesos adaptativos, y del papel de la selección natural en la adaptación a nivel molecular. En particular, se ha estudiado el origen y evolución de las familias multigénicas del sistema quimiosensorial en artrópodos, y el proceso de adaptación trófica que ha tenido lugar durante la radiación adaptativa de las arañas del género *Dysdera* en las Islas Canarias.

1 Desarrollo e implementación de nuevos métodos para el estudio de familias multigénicas en ensamblajes genómicos

La disponibilidad actual de secuencias genómicas completas en especies no modelo supone una oportunidad sin precedentes para estudiar el proceso de adaptación a nivel molecular[149]. No obstante, los datos generados con técnicas de secuenciación masiva presentan una serie de inconvenientes que dificultan su uso en primera instancia[156,157]. Una de los principales limitaciones reside en la anotación de las secuencias genómicas una vez ensambladas, especialmente en organismos "exóticos" que carecen de modelos génicos preexistentes (organismos no modelo)[158,159]. Existen métodos, como los implementados en MAKER2 o BRAKER1, que combinan evidencias de diversas fuentes (predicciones *ab initio*, RNA-seq y modelos génicos definidos en otras especies) para llevar a cabo la anotación estructural[159,160]. Sin embargo, en muchos modelos génicos, estas anotaciones automáticas distan de ser precisas, siendo especialmente relevante en el caso de las familias multigénicas[161,162]. Las familias de genes presentan una serie de características que dificultan considerablemente la correcta anotación de sus miembros en ensamblajes genómicos. De forma general, las nuevas copias son producto de duplicaciones génicas originadas por entrecruzamiento desigual y se encuentran localizadas en formaciones de genes en tándem en la misma región genómica[102,163]. Esta configuración ocasiona frecuentemente ensamblajes erróneos y anotaciones incorrectas de los genes, produciendo fusiones génicas, genes incompletos o genes quimera, o incluso impidiendo por completo la generación de modelos génicos en la región. Además, algunos miembros de la familia pueden presentar una alta divergencia a nivel de secuencia, lo que puede dificultar su identificación con métodos clásicos basados en similitud de secuencia, requiriendo el uso de metodologías y herramientas de identificación de homólogos remotos más potentes[100,164].

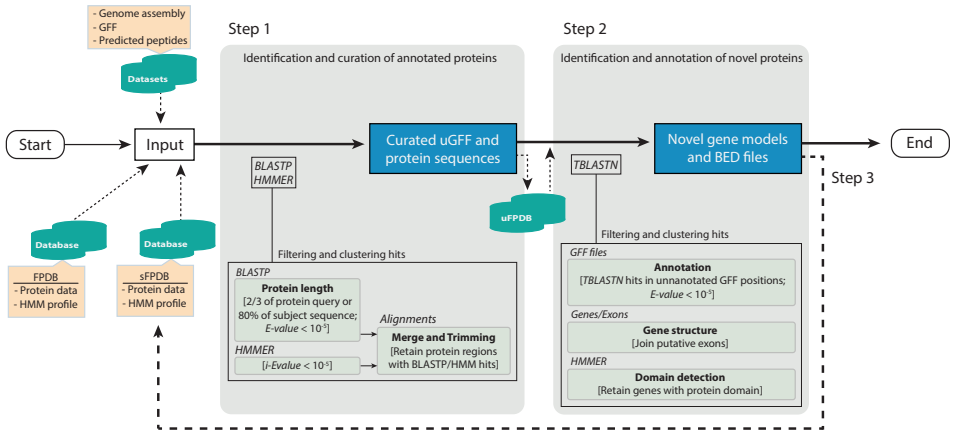


Figura 9. Flujo de trabajo de BITACORA.

En este contexto, hemos desarrollado la aplicación bioinformática BITACORA para facilitar la identificación y correcta anotación estructural de miembros de familias multigénicas en ensamblajes genómicos. En esta aplicación se combinan herramientas de búsqueda de genes homólogos por metodologías basadas en similitud de secuencia, como BLAST[165] y HMMER (este último basado en perfiles probabilísticos)[166], además de una serie de scripts escritos en el lenguaje de programación Perl para conectar y automatizar las múltiples tareas implementadas (Figura 9). BITACORA permite no sólo identificar aquellos genes pertenecientes a una familia multigénica focal que se encuentran ya anotados en el genoma a estudiar, y corregir, si es necesario, el modelo predicho (véanse fusiones génicas, quimeras…), sino que también puede identificar y generar nuevos modelos para los miembros de la familia que no habían sido anotados previamente. El uso de BITACORA en el estudio comparativo de familias multigénicas en diferentes genomas es fundamental para la rigurosidad de posteriores análisis, ya que la correcta identificación y anotación de todos los miembros de una familia es crítico a la hora de estimar las tasas de ganancia y pérdida de genes en distintos linajes, determinar la presencia o ausencia de una familia multigénica en un grupo determinado de organismos, o detectar la huella de la selección natural en sus miembros.

2 Origen y evolución de las familias multigénicas del sistema quimiosensorial en artrópodos

El SQ es un sistema crítico para la supervivencia y la reproducción de prácticamente todos los organismos vivos, ya que participa en funciones esenciales como la detección de alimento, defensa ante depredadores, búsqueda de pareja y cortejo e interacciones sociales[36,37]. Su papel específico en actividades reproductivas podría incluso contribuir en procesos evolutivamente relevantes como el aislamiento reproductivo y la especiación[36,167]. En términos generales, el SQ comprende tanto el gusto (detección de compuestos solubles), como el olfato (compuestos volátiles). En artrópodos, la diversificación de los distintos subfilos tuvo lugar con anterioridad a la colonización del medio terrestre[116]. Por lo tanto, las estrategias para adaptarse a la detección y reconocimiento de señales químicas en el medio aéreo tienen que haberse originado de forma independiente, como mínimo, en los grandes subfilos de artrópodos (Figura 3). La disponibilidad actual de genomas completos de quelicerados, junto con el acceso a nuevas tecnologías de secuenciación debido a sus reducidos costes, nos ha permitido analizar el origen y evolución de las principales familias multigénicas del SQ a lo largo del filo de los artrópodos, especialmente en los quelicerados. No obstante, algunos de los genomas disponibles en las bases de datos están altamente fragmentados (la continuidad de las secuencias no es la deseada) y presentan anotaciones estructurales y funcionales que distan de ser completas y/o correctas. En consecuencia, y para poder realizar un análisis riguroso de la evolución de las familias del SQ en quelicerados, desarrollamos y utilizamos la herramienta BITACORA para anotar e identificar los miembros de estas familias en dichos genomas (Figura 10).

2.1 Quimiorreceptores en artrópodos

En este trabajo, hemos identificado miembros de las familias de IRs/iGluRs y GRs en todos los genomas de quelicerados estudiados (Figura 10); por el contrario, no se ha detectado ningún gen o secuencia parcial relacionada con la familia de los ORs, en conformidad con lo reportado en los análisis previos a esta tesis[71–74,168]. Consecuentemente, también hemos encontrado transcritos de las dos primeras familias de receptores, tanto específicos como con expresión diferencial, en el transcriptoma de los apéndices quimiosensoriales (primer par de patas y palpos) de la araña endémica de las Islas Canarias *D. silvatica*. Sorprendentemente, los repertorios (número e identidad a nivel de secuencia de las copias de cada familia)

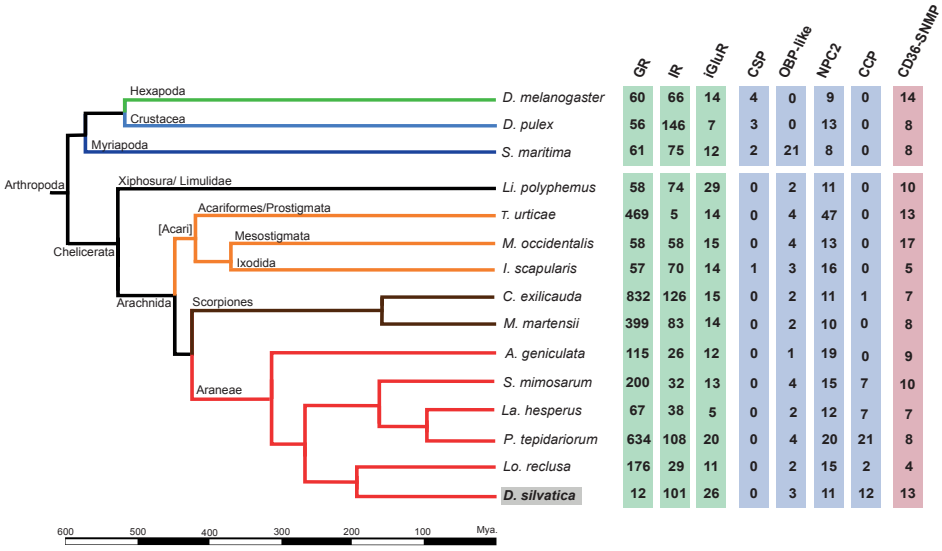


Figura 10. Relaciones filogenéticas entre las 15 especies estudiadas en esta tesis. Los tiempos de divergencia se indican en millones de años. Cada linaje principal está representado por un color: verde, insectos; azul claro, crustáceos; azul oscuro, miriápodos; negro, xifosuros; naranja, ácaros; marrón, escorpiones; y rojo, arañas. En la parte derecha se muestra el número mínimo de miembros de cada familia multigénica estimado a partir de las secuencias genómicas, con la excepción de *D. silvatica* donde corresponden a transcritos.

de estas familias varia ampliamente entre los distintos linajes de quelicerados, con enormes diferencias acumuladas en linajes particulares, un patrón similar a lo descrito en algunas especies de insectos[66]. De hecho, las tasas de nacimiento y muerte que hemos estimado para quelicerados son similares a las descritas en el género *Drosophila*[169]. Se ha sugerido que el número de quimiorreceptores en un genoma particular podría correlacionar con algunos aspectos ecológicos complejos en los artrópodos, como por ejemplo una dieta ampliamente generalista[168,170]. De hecho, hemos observado que en especies generalistas de quelicerados, como el ácaro *T. urticae*, la araña común *P. tepidariorum*, o los dos escorpiones incluidos en nuestros análisis, existe un mayor número de receptores, por lo que nuestros resultados podrían estar apoyando dicha hipótesis.

Los análisis filogenéticos de las diferentes copias de una misma familia muestran que los quelicerados poseen su propio repertorio específico de GRs, no detectándose de hecho ningún clado que incluya proteínas de distintos subfilos de artrópodos. Notablemente, los miembros de esta familia conservados en insectos y crustáceos, y que se han caracterizado funcionalmente como receptores de azúcares o CO_2 en

varias especies[65,68], no se encuentran (o al menos no existen secuencias similares) en los genomas de quelicerados. La mayoría de clados filogenéticos son incluso especie-específicos, soportando un modelo de evolución por nacimiento y muerte extraordinariamente dinámico para esta antigua familia de receptores. No obstante, sí que encontramos un clado monofilético con representantes de todos los linajes de quelicerados estudiados, lo cual sugeriría la existencia de receptores con una función más conservada dentro de este grupo. Algunos de estos receptores potencialmente gustativos se encuentran expresados de forma específica en los palpos y las patas de la araña *D. silvatica*, así como en tejidos quimiosensoriales de ácaros[161,171], indicando que podrían tener una función relacionada con el SQ. Nuestros datos corroboran por tanto que los GRs con función quimiosensorial se originaron de forma previa a la diversificación de los artrópodos, proteínas que podrían denominarse como GRs ancestrales (aGRs), a partir de los GRLs (Figura 11). Los aGRs diversificaron posteriormente para formar los actuales GRs, los cuales habrían evolucionado de forma independiente en los distintos subfilos de artrópodos por un mecanismo de nacimiento y muerte muy dinámico, fenómeno que podría haber estado asociado en algún caso a aspectos ecológicos y de la dieta en quelicerados.

En el caso de los IRs, el patrón filogenético observado es, de forma general, similar al de los GRs, siendo la mayoría de clados especie-específicos. En esta subfamilia, sin embargo, sí existen miembros conservados en todos los artrópodos (IR8a, IR25a, IR76b e IR93a presentan clados monofiléticos formados por copias identificadas en genomas de diferentes filos). De hecho, la caracterización del receptor IR8a en el cangrejo de herradura *Limulus polyphemus* (Xiphosura), anotación basada en el árbol de esta familia y que presenta un buen soporte filogenético, implicaría reformular la hipótesis propuesta por Eyun y colaboradores[41], indicando que este gen ya estaba presente en el ancestro de artrópodos y que, por lo tanto, se perdió en el ancestro de los arácnidos, algo que además soportaría la monofilia de Arachnida (también en duda; Figura 11)[44,123]. No obstante, no hemos encontrado ninguna copia filogenéticamente relacionada con los IR21a e IR40a en los genomas analizados de quelicerados, lo que sugiere que los genes que habíamos descrito previamente como posibles homólogos de estos dos IR en el transcriptoma de *Dysdera silvatica*[164], eran producto de artefactos en la reconstrucción filogenética a partir de alineamientos de baja calidad incluyendo transcritos parciales de una familia tan divergente[172]. Además de utilizar secuencias genómicas completas, para el análisis comparativo de los 11 quelicerados usamos un programa de reconstrucción filogenética, PaHMM-Tree, que no requiere un alineamiento previo para obtener las filogenias de las familias multigénicas del SQ[162,172]. Por otro lado, también detectamos expresión de algunos miembros de la familia de los IRs en las patas y palpos de *D. silvatica*,

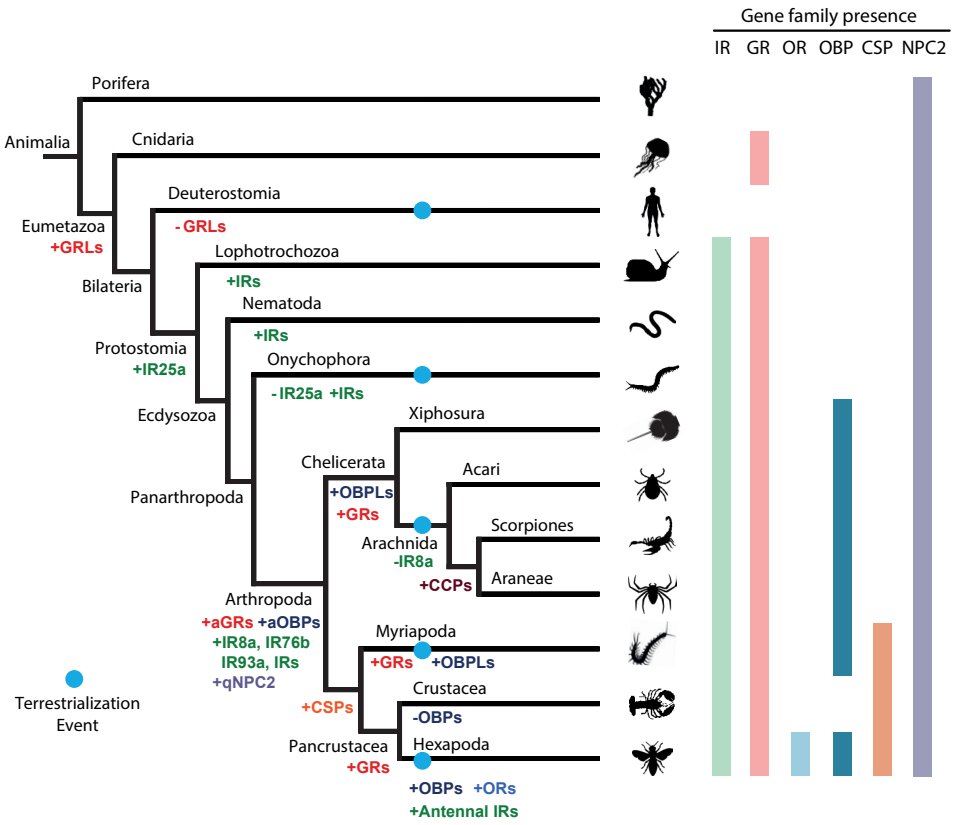


Figura 11. Historia evolutiva de las principales familias del SQ en artrópodos (adaptación de la Figura 3 incorporando los resultados obtenidos en esta tesis). La presencia o ausencia de las distintas familias multigénicas en cada linaje se indica en columnas (parte de la derecha). Sobre las ramas se representa la aparición (+) o pérdida (-) de una familia o alguno de sus miembros[41–43]. IR: *Ionotropic receptor*; GR: *Gustatory receptor*; OR: *Odorant receptor*; OBP: *Odorant binding protein*; CSP: *Chemosensory protein*; NPC2: *Niemann-Pick C2 protein*. Los puntos azules denotan colonizaciones independientes del medio terrestre en cada linaje. La relación filogenética entre los quelicerados estudiados es la más soportada actualmente, y asume la monofilia de arácnidos[44].

destacando la expresión diferencial en el primer par de patas de esta araña de IR25a, propuesto como correceptor de otros IR en la antena de *Drosophila*[87]. En conjunto, todas las evidencias apuntan a la presencia de IRs con función quimiosensorial en el genoma del ancestro de artrópodos, y que esta familia ha evolucionado a través de un modelo de nacimiento y muerte durante su diversificación (aunque menos dinámico que el observado en las GRs).

En conclusión, hemos demostrado que los GRs e IRs siguen una dinámica de evolución por nacimiento y muerte, con una alta tasa de ganancia y pérdida de genes. En particular, expansiones y contracciones ocurridas aparentemente de forma episódica e, incluso probablemente específica de linajes particulares, habrían contribuido a generar la enorme diferencia entre repertorios observados entre los quelicerados (no sólo en cuanto al número de copias sino también a la similitud de secuencia de las proteínas), algo ya sugerido anteriormente por otros autores[72,173]. Nuestros datos confirman la presencia de estas dos familias multigénicas en el ancestro de los artrópodos, como también había sido previamente sugerido[41,42], y que la función de algunos de sus miembros estaría ya relacionada con la quimiopercepción. Por lo tanto, los distintos linajes de artrópodos habrían reclutado (mediante divergencia a nivel de secuencia o cambios en los patrones de expresión) genes ya existentes (que realizarían una función similar) de forma independiente para adaptarse a la detección de estímulos químicos en el medio terrestre. Finalmente, nuestros análisis también confirman la presencia en quelicerados de otras familias multigénicas del SQ de insectos, como las PPKs y SNMPs. Estos genes, que a diferencia de los receptores anteriores no tienen una función principal asociada a la quimiopercepción en los animales, estarían también presentes en el ancestro de los artrópodos y podrían haber sido cooptadas por los distintos linajes para la quimiopercepción, como ha sido sugerido en las PPKs en el ácaro *T. urticae* dónde el repertorio de IRs es limitado[168].

2.2 Proteínas solubles secretadas

Previo al inicio de esta tesis doctoral, las OBPs se habían caracterizado únicamente en insectos mientras que las CSPs se habían detectado en especies de todos los subfilos de artrópodos[100,106]. Además, su similitud estructural remota había llevado a postular que las OBPs surgieron a partir de CSPs en el ancestro de los insectos[100]. No obstante, en quelicerados, las CSPs se habían encontrado únicamente en el genoma de *Ixodes scapularis*[73,174]. En la misma línea, nosotros no hemos encontrado evidencias de la presencia de CSPs en ninguno de los genomas de quelicerados, incluyendo aquellos en los que Eyun y colaboradores[41] reportaron la presencia de unas secuencias similares a las CSPs, pero que contenían codones *stop* y carecían del patrón característico de cisteínas de esta familia (Figura 10). Nuestros resultados sugieren por lo tanto que la presencia de CSPs en este subfilo es completamente cuestionable. De hecho, recientemente, se ha constatado que la secuencia descrita como CSP en *I. scapularis* es idéntica a una CSP del mosquito *C. quinquefasciatus*,

lo que sugeriría una contaminación y reforzaría la hipótesis de la ausencia total de CSPs en quelicerados[175].

Por otro lado, nuestro estudio ha producido un resultado muy sorprendente, la presencia de OBPs en quelicerados y miriápodos, proteínas a las que hemos denominado OBP-*like* (OBPL). Cabe destacar que, durante el transcurso de esta tesis, Renthal y colaboradores[176] también obtuvieron el mismo resultado en un estudio totalmente independiente. Las OBPL están conservadas en todos los quelicerados, encontrándose entre 1 y 4 miembros por especie (Figura 10). Dado el bajo número de copias y su expresión generalizada en todos los tejidos del transcriptoma de *D. silvatica*, su función podría no estar relacionada con la quimiopercepción en este grupo. De hecho, en insectos, además de su función en el SQ, también se ha determinado que las OBPs participan como moléculas transportadoras en otros procesos biológicos[109]. Sin embargo, otros estudios sí han encontrado OBPLs expresadas en órganos quimiosensoriales en ácaros[161,175,176], por lo que no se descarta totalmente su participación en la quimiopercepción en algunas especies de quelicerados.

Las NPC2 han sido propuestas por Pelosi y colaboradores[106] como posibles proteínas solubles del SQ en algunas especies de artrópodos. Esta familia se encuentra presente en quelicerados con un repertorio que varía ligeramente entre los distintos linajes (entre 10 y 20 copias, con la excepción de las 47 observadas en *T. urticae*; Figura 10). Estudios recientes han reportado la expresión de sus miembros en los apéndices quimiosensoriales en distintas especies de quelicerados, especialmente en ácaros, además de en insectos, apoyando así a la hipótesis de su función como proteínas transportadoras de odorantes[161,175,176]. No obstante, en el transcriptoma de *D. silvatica* sólo encontramos un miembro (de los 11 expresados) con expresión específica en el primer par de patas. La función de estas proteínas en quelicerados dista de ser comprendida, aunque las evidencias indicarían que son buenos candidatos como proteínas solubles del SQ de quelicerados.

Un aspecto relevante de esta tesis doctoral es la identificación de una nueva familia de proteínas no descritas y sin similitud con ninguna secuencia conocida, a la que hemos denominado como *candidate carrier proteins* (CCPs). Las CCPs son proteínas probablemente solubles y secretadas (presentan una estructura globular con aminoácidos altamente hidrofóbicos en la superficie y un claro péptido señal) que presentan una similitud estructural remota con las OBPs de insectos, y se encuentran expresadas en los apéndices quimiosensoriales de la araña *D. silvatica*. Una vez caracterizadas en *D. silvatica*, hemos detectado miembros esta familia

multigénica en todas las especies de arañas estudiadas en esta tesis (Figura 10). Los miembros de esta familia representarían un nuevo candidato para llevar a cabo la función relacionada con la detección y solubilización de estímulos químicos que las OBPs y CSPs efectúan en insectos, aunque su función específica, como en el caso de las NPC2, aún debe ser estudiada en profundidad.

En conclusión, nuestros datos, integrados con el conocimiento proporcionado por otros trabajos publicados durante el transcurso de esta tesis, requieren una reformulación de las ideas sobre el origen de las OBPs y CSPs, y en general, de las proteínas solubles secretadas en artrópodos. Así, en el ancestro de los artrópodos existiría el predecesor de las OBPs (denominado aOBP) a partir del cual surgirían las OBPs en insectos y las OBPL en quelicerados y miriápodos, perdiéndose esta familia en los crustáceos (Figura 11). A su vez, es posible que las CSPs se originasen en el ancestro de los mandibulados (linaje que comprende los miriápodos y pancrustáceos) a partir de las aOBPs, aunque no hay evidencias concluyentes sobre este aspecto. En el caso de las CCPs, se originarían en el ancestro de arañas y escorpiones, pero se desconoce su origen concreto dada la ausencia de estas proteínas en otros linajes de artrópodos. Finalmente, las NPC2 habrían expandido su repertorio y adquirido su posible función quimiosensorial en el ancestro de los artrópodos (denominadas qNPC2; Figura 11).

3 Determinantes genómicos de la especialización trófica convergente en *Dysdera*

Las arañas del género *Dysdera* presentan una de las radiaciones insulares más espectaculares entre los arácnidos[127]. Unas 47 especies de éste género han sido catalogadas como endémicas de las Islas Canarias, varias de ellas presentando una especialización trófica en la alimentación a base de isópodos[126,127]. Esta especialización de la dieta (estenofagia) se ha originado en múltiples ocasiones de forma independiente en el género *Dysdera*, tanto en las islas como en el continente, y está asociada a modificaciones en los quelíceros de las arañas especialistas, que a su vez se relacionan con diferentes estrategias en la captura de los isópodos. La convergencia fenotípica observada durante la diversificación insular de estas arañas ofrece un buen modelo para estudiar el papel de la selección natural en las radiaciones adaptativas y aportar información relevante en el debate sobre la predictibilidad de los procesos evolutivos.

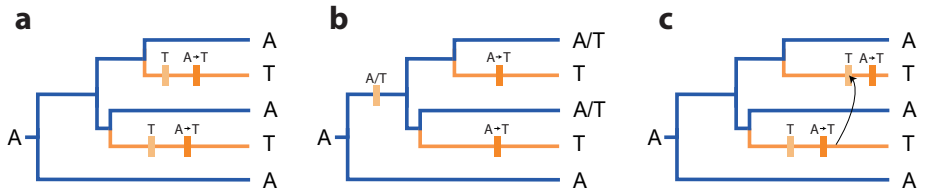


Figura 12. Ilustración en un contexto filogenético de los posibles mecanismos que pueden generar cambios convergentes a nivel molecular: **a**) Evolución paralela; **b**) Fijación colateral de variación ancestral compartida; **c**) Introgresión. El color de la rama indica convergencia fenotípica (azul, generalistas; naranja, especialistas). Adaptado de Stern[32].

En esta tesis, hemos secuenciado y comparado los transcriptomas de cinco especies de *Dysdera* endémicas de las Islas Canarias, incluyendo dos parejas de especies hermanas, una de ellas especialista y la otra generalista, que representan con casi toda seguridad dos eventos independientes de adaptación a la estenofagia (Figura 8). Nuestros resultados demuestran que, al menos en parte, la especialización se ha producido a través de cambios en el mismo gen (en ciertos casos incluso en el mismo aminoácido) o en genes con funciones muy similares, tanto en sus regiones reguladoras como codificadoras. Cabe destacar que el cambio en el número de copias de la misma familia multigénica podría haber tenido también un papel importante en esta adaptación. No obstante, el proceso de ensamblaje *de novo* de los transcritos a partir de secuencias cortas obtenidas por NGS (*reads*; 100 pares de bases) es muy problemático, especialmente en familias multigénicas, impidiendo así la determinación robusta del número real de copias y de su expresión. Por tanto, la naturaleza de los datos nos imposibilitó realizar inferencias sólidas sobre posibles cambios significativos en el repertorio de familias multigénicas (como por ejemplo las implicadas en el SQ) asociados a la estenofagia. Sin embargo, la publicación reciente del ensamblaje genómico de *Dysdera silvatica*[177] nos permitirá abordar en un futuro cercano el estudio de las familias de quimiorreceptores y proteínas solubles secretadas bajo esta hipótesis de estudio.

En este trabajo hemos identificado variantes genéticas compartidas únicamente entre las especies especialistas, tanto a nivel de cambios en la región codificadora como de expresión diferencial, y que son concordantes con la convergencia fenotípica observada. Estas variantes pueden tener distintos orígenes (Figura 12)[32]. Por un lado, los cambios genéticos podrían deberse a nuevas mutaciones que se habrían originado de forma independiente en los dos linajes especialistas (evolución paralela que genera patrones filogenéticos de homoplasia; Figura 12a). Los cambios observados también podrían deberse a la fijación diferencial de variación genética presente en el ancestro de estas especies, o a la introgresión de variantes desde una

especie especialista a la otra debido a procesos de hibridación (Figura 12bc). No obstante, en los tres casos el proceso podría ocurrir de forma neutral (por azar) o por la acción de la selección natural positiva. En el primer caso, nuestros resultados indican que el número de mutaciones potencialmente convergentes es superior al que se esperaría en una evolución paralela simplemente por azar. En el segundo y tercer caso, hemos evaluado la probabilidad de fijación únicamente por azar de variantes genéticas compartidas en el ancestro de las especies especialistas que pudieran generar falsos de patrones homoplasia (hemiplasia), o de introgresión entre nuestras especies, resultando ambos altamente improbables. A pesar de que observamos un porcentaje importante de árboles incongruentes (árboles de genes diferentes al árbol de especies; las cuatro especies de *Dysdera* estudiadas han divergido recientemente y aún existe un alto porcentaje de polimorfismo compartido entre ellas), estas incongruencias raramente emparejan las dos especies especialistas con un ancestro común más reciente que con las otras generalistas (lo que podría indicar hemiplasia en esos genes), y nunca en los genes candidatos. Así, los resultados indican que la fijación de polimorfismos compartidos o la introgresión (tanto neutral como bajo selección) no explicaría los cambios convergentes observados en especialistas. Como conclusión, podemos decir que la explicación más plausible para la convergencia molecular observada en los genes candidatos sería la de los cambios paralelos independientes en cada uno de los linajes especialistas guiados por la selección positiva.

No obstante, en el caso de los candidatos de expresión diferencial, esta aproximación podría ser incorrecta dado que la región del transcrito secuenciada y la región responsable de la regulación de la expresión del gen podrían no estar ligadas (ligamiento parcial entre el gen y la región reguladora en *cis*, o incluso debido a la regulación en *trans* por elementos reguladores localizados en regiones no ligadas al gen). Esto ocasionaría que los árboles de genes inferidos a partir de las regiones codificadoras no fueran informativos acerca de una posible incongruencia indicativa de hemiplasia. No obstante, como hemos detallado anteriormente, la probabilidad de hemiplasia afectando a las dos especialistas es muy baja en general según nuestras estimaciones. Además, las funciones asociadas a los genes candidatos están enriquecidas en procesos biológicos relevantes para el proceso de especialización trófica estudiado. En consecuencia, nuestros resultados indican que el modelo de estudio es robusto y nos permite identificar cambios concordantes con la convergencia fenotípica observada a nivel molecular, y que estarían implicados en el evento de especialización trófica en *Dysdera*. Los procesos biológicos en los que están implicados los genes candidatos son principalmente la detoxificación y homeostasis de metales pesados, ya sea de forma directa en funciones de unión y transporte

de metales pesados, o a través de vías secundarias relacionadas con su toxicidad, como el daño oxidativo por la formación de especies reactivas del oxígeno (ROS) o el efecto sobre la actividad del sistema inmune. A su vez, también encontramos genes relacionados con el metabolismo de nutrientes esenciales, además de algunas toxinas que podrían tener un papel adaptativo en el evento de especialización trófica estudiado. Sin embargo, nuestro trabajo es un punto de partida para el estudio de la radiación adaptativa de *Dysdera* a nivel molecular. De hecho, el análisis de un mayor número de especies especialistas sería necesario para validar nuestros resultados y, en general, el papel de la evolución paralela en la generación de nuevos cambios concordantes con la convergencia fenotípica. A su vez, la disponibilidad actual de datos genómicos en *D. silvatica* nos permitirá estudiar el efecto de la selección natural positiva en la región reguladora de los candidatos de expresión diferencial.

En resumen, nuestros resultados demuestran que la colonización de las Islas Canarias ha ofrecido una oportunidad ecológica única a las arañas del género *Dysdera*, favoreciendo una radiación adaptativa en la que la diversificación de las especies ha tenido lugar de forma concomitante junto con eventos repetidos de especialización trófica. Este modelo nos proporciona un ejemplo excepcional sobre la repetitividad en la evolución y en el estudio de los mecanismos genómicos implicados en la convergencia fenotípica en estos organismos. De hecho, los resultados obtenidos apoyan la idea de que la convergencia fenotípica puede ser producida por variantes genéticas que han ocurrido de forma repetida y a distintos niveles jerárquicos a lo largo de la evolución[178], y que en el caso de *Dysdera* ha implicado cambios convergentes en el mismo aminoácido, en el mismo gen o en genes con funciones equivalentes. También demostramos que la selección natural ha promovido la fijación de algunos de estos cambios, confirmando el papel de las fuerzas adaptativas como un determinante fundamental en la convergencia fenotípica[179]. Por tanto, nuestro estudio no solo proporciona conocimiento relevante sobre la base genómica de la adaptación, sino que también ofrece nuevas perspectivas en el debate acerca de la predictibilidad de la evolución a nivel molecular.

Conclusiones

1. En esta tesis se ha desarrollado BITACORA, una herramienta bioinformática para la identificación y anotación de familias multigénicas en ensamblajes genómicos o transcriptómicos de organismos no modelo.

2. BITACORA ha permitido identificar miles de nuevas copias de familias multigénicas del sistema quimiosensoriral en los genomas de quelicerados, así como subsanar modelos génicos erróneos en las anotaciones existentes.

3. Las familias multigénicas de las GR y las IR codifican los principales quimiorreceptores que median la respuesta a estímulos químicos en quelicerados, y, por lo tanto, de forma general en todos los artrópodos.

4. Las familias de los GR e IR han evolucionado bajo un proceso muy dinámico de ganancia y pérdida de genes con la presencia de expansiones y contracciones episódicas en linajes específicos de quelicerados.

5. Los receptores IR8a, IR25a, IR76b e IR93a son los únicos IRs conservados en todos los subfilos de artrópodos, destacando la pérdida de IR8a en el grupo de los arácnidos.

6. Los subfilos de quelicerados y miriápodos codifican una familia de proteínas relacionadas con las OBPs de insectos, las OBP-*like*, sugiriendo un origen de esta familia de proteínas solubles más antiguo del que había sido reportado, y que se puede trazar previo a la diversificación de los principales linajes de artrópodos.

7. La familia de los CSP se encuentra ausente en quelicerados, confirmando así que el origen de esta familia se puede trazar al ancestro de los mandibulados.

8. Los genomas de arañas codifican una nueva familia multigénica de proteínas solubles hasta ahora desconocida, denominada como CCPs, cuya expresión específica en los apéndices quimiosensoriales de *D. silvatica* sugiere un posible rol en la quimiopercepción.

9. La familia multigénica de las NPC2 se encuentra expandida en artrópodos con respecto al repertorio observado en vertebrados, y podría estar implicada en la unión y transporte de estímulos químicos.

10. Los cambios genéticos asociados a la convergencia fenotípica (especialización trófica) observada en el género *Dysdera* se han producido a distintos niveles jerárquicos (mismo aminoácido, gen o función génica).

11. Algunas de las variantes que se han acumulado de forma paralela en los linajes especialistas han evolucionado por selección positiva, confirmando el papel de las fuerzas adaptativas como uno de los principales determinantes en la convergencia evolutiva.

12. Los cambios genéticos convergentes asociados a la especialización trófica en *Dysdera* están relacionados con la secreción y detoxificación de metales pesados acumulados en las presas de las arañas especialistas (isópodos terrestres), aunque también han participado otros cambios, como aquellos en la asimilación de nutrientes o componentes del veneno.

13. En términos generales, nuestro estudio demuestra que la evolución, especialmente en el contexto de la adaptación, puede actuar de forma repetida y, por tanto, presenta un componente predecible. Por un lado, los distintos subfilos de artrópodos han cooptado las mismas familias multigénicas del sistema quimiosensorial durante su adaptación al medio terrestre, lo cual pone de manifiesto la importancia de la variación genética ancestral en el proceso adaptativo. Por otro lado, demostramos el papel fundamental de la evolución paralela en la radiación adaptativa del género *Dysdera* en Canarias.

Conclusions

1. We have developed BITACORA, a bioinformatics tool to facilitate the identification and annotation of gene families in genome or transcriptome assemblies of non-model organisms.

2. BITACORA allowed the identification of thousands of new chemosensory gene family copies in chelicerate genomes and the curation of many gene models in the existing annotations.

3. The GR and IR gene families encode the chemoreceptors that mediate the response to chemical stimuli in chelicerates and, in general, in arthropods.

4. The GR and IR families have evolved under a very dynamic gene birth and death process, influenced by episodic bursts of gene duplication yielding lineage-specific expansions.

5. The receptors IR8a, IR25a, IR76b and IR93a are the only IRs conserved across arthropods, although the IR8a has been lost in arachnids.

6. Chelicerates and myriapods subphylum encode a family of proteins distantly related to the insect OBPs, the OBP-*like*, suggesting a much older origin than previously thought for this family of soluble proteins, and that can be traced back prior to the diversification of the major arthropod linages.

7. The CSP family is completely absent in chelicerates, confirming that the origin of this family can be traced back to the ancestor of Mandibulata (i.e., hexapods and crustaceans).

8. The genomes of spider species encode a novel (previously uncharacterized) gene family of soluble proteins with members expressed in the chemosensory appendages of *D. silvatica*, named as CCPs, which would suggest an active role in spider quimioperception.

9. The NPC2 gene family has expanded in arthropods (relative to the conserved repertoire observed in vertebrates), where it could be involved in the binding and transport of chemical cues.

10. The phenotypic convergence (accompanying dietary specializations) observed in the genus *Dysdera* involved repeated genetic changes at different hierarchical levels (in particular positions, genes or gene functions).

11. Some of the genetic variants that have been repeatedly accumulated in specialist lineages are promoted by positive selection, supporting the view that adaptive forces are primary determinants of evolutionary convergence.

12. Most of the targets of convergent changes associated with the trophic specialization in *Dysdera* are related to the excretion and detoxification of heavy metals accumulated in the preferred prey of specialist spiders (terrestrial isopods), the assimilation of nutrients and venom components.

13. Globally, our study demonstrate that adaptive evolution shows repeatability and, therefore, presents a predictable component. On one hand, the different arthropods subphylum have coopted the same chemosensory gene families during the adaptation to the terrestrial environment, which proves the importance of ancestral genetic variation in the adaptive process. On the other hand, we found that parallel evolution had a crucial role in the adaptive radiation of the spider genus *Dysdera* in the Canary Islands.

Bibliografía

1. Darwin, C. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. (John Murray, 1859).
2. Hartl, D. L. & Clark, A. G. *Principles of population genetics*. (Sinauer Associates, 2007).
3. Kimura, M. Evolutionary Rate at the Molecular Level. *Nature* **217**, 624–626 (1968).
4. King, J. L. & Jukes, T. H. Non-Darwinian evolution. *Science* **164**, 788–98 (1969).
5. Kimura, M. *The neutral theory of molecular evolution*. (Cambridge University Press, 1983).
6. Ohta, T. Slightly deleterious mutant substitutions in evolution. *Nature* **246**, 96–98 (1973).
7. Crow, J. F. & Kimura, M. *An introduction to population genetics theory*. (Harper & Row, 1970).
8. Barton, N. H., Briggs, D. E. G., Eisen, J. A., Goldstein, D. D. & Patel, N. H. *Evolution*. (Cold Spring Harbor Laboratory Press, 2007).
9. Calvo-Martín, J. M. *Evolución molecular de los genes del grupo Polycomp en el género Drosophila* [tesis doctoral]. (Universitat de Barcelona, 2017).
10. Miyata, T. & Yasunaga, T. Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J. Mol. Evol.* **16**, 23–36 (1980).
11. Almeida, F. C., Sánchez-Gracia, A., Walden, K. K. O., Robertson, H. M. & Rozas, J. Positive selection in extra cellular domains in the diversification of *Strigamia maritima* chemoreceptors. *Front. Ecol. Evol.* **3**, 1–9 (2015).
12. Torres-Oliva, M., Almeida, F. C., Sánchez-Gracia, A. & Rozas, J. Comparative Genomics Uncovers Unique Gene Turnover and Evolutionary Rates in a Gene Family Involved in the Detection of Insect Cuticular Pheromones. *Genome Biol. Evol.* **8**, 1734–1747 (2016).
13. Murrell, B. *et al.* Detecting Individual Sites Subject to Episodic Diversifying Selection. *PLoS Genet.* **8**, e1002764 (2012).
14. Yang, Z. & Swanson, W. J. Codon-Substitution Models to Detect Adaptive Evolution that Account for Heterogeneous Selective Pressures Among Site Classes. *Mol. Biol. Evol.* **19**, 49–57 (2002).
15. Smith, M. D. *et al.* Less Is More: An Adaptive Branch-Site Random Effects Model for Efficient Detection of Episodic Diversifying Selection. *Mol. Biol. Evol.* **32**, 1342–1353 (2015).
16. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–95 (1989).
17. Hudson, R. R., Kreitman, M. & Aguadé, M. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153–9 (1987).
18. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**, 652–654 (1991).

19. Mergeay, J. & Santamaria, L. Evolution and Biodiversity: the evolutionary basis of biodiversity and its potential for adaptation to global change. *Evol. Appl.* **5**, 103–106 (2012).
20. MacArthur, R. H. & Wilson, E. O. *The Theory of Island Biogeography*. (Princeton University Press, 1967).
21. Mayr, E. *Systematics and the Origins of Species*. (Columbia University Press, 1942).
22. Whittaker, R. J. & Fernández-Palacios, J. M. *Island biogeography: ecology, evolution, and conservation*. (Oxford University Press, 2007).
23. Almén, M. S. *et al.* Adaptive radiation of Darwin's finches revisited using whole genome sequencing. *BioEssays* **38**, 14–20 (2016).
24. Muschick, M., Indermaur, A. & Salzburger, W. Convergent Evolution within an Adaptive Radiation of Cichlid Fishes. *Curr. Biol.* **22**, 2362–2368 (2012).
25. Gillespie, R. G. & Roderick, G. K. Arthropods on Islands: Colonization, Speciation, and Conservation. *Annu. Rev. Entomol.* **47**, 595–632 (2002).
26. Losos, J. B. & Ricklefs, R. E. Adaptation and diversification on islands. *Nature* **457**, 830–836 (2009).
27. Rosenblum, E. B., Parent, C. E. & Brandt, E. E. The Molecular Basis of Phenotypic Convergence. *Annu. Rev. Ecol. Evol. Syst.* **45**, 203–226 (2014).
28. Shapiro, M. D. *et al.* Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* **428**, 717–723 (2004).
29. Albert, A. Y. K. *et al.* The genetics of adaptive shape shift in stickleback: pleiotropy and effect size. *Evolution* **62**, 76–85 (2008).
30. Chan, Y. F. *et al.* Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science* **327**, 302–5 (2010).
31. Jones, F. C. *et al.* The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55–61 (2012).
32. Stern, D. L. The genetic causes of convergent evolution. *Nat. Rev. Genet.* **14**, 751–764 (2013).
33. Marques, D. A., Meier, J. I. & Seehausen, O. A Combinatorial View on Speciation and Adaptive Radiation. *Trends Ecol. Evol.* (2019). doi:10.1016/j.tree.2019.02.008
34. Schluter, D. & Conte, G. L. Genetics and ecological speciation. *Proc. Natl. Acad. Sci. USA* **106 Suppl 1**, 9955–62 (2009).
35. Van Belleghem, S. M. *et al.* Evolution at two time frames: Polymorphisms from an ancient singular divergence event fuel contemporary parallel evolution. *PLOS Genet.* **14**, e1007796 (2018).
36. Smadja, C. & Butlin, R. K. On the scent of speciation: the chemosensory system and its role in premating isolation. *Heredity* **102**, 77–97 (2009).

37. Asahina, K., Pavlenkovich, V. & Vosshall, L. B. The survival advantage of olfaction in a competitive environment. *Curr. Biol.* **18**, 1153–5 (2008).

38. Stensmyr, M. C. *et al.* Insect-Like Olfactory Adaptations in the Terrestrial Giant Robber Crab. *Curr. Biol.* **15**, 116–121 (2005).

39. Krang, A.-S., Knaden, M., Steck, K. & Hansson, B. S. Transition from sea to land: olfactory function and constraints in the terrestrial hermit crab *Coenobita clypeatus*. *Proc. R. Soc. B Biol. Sci.* **279**, 3510–3519 (2012).

40. McClintock, T. S. & Ache, B. W. Ionic currents and ion channels of lobster olfactory receptor neurons. *J. Gen. Physiol.* **94**, 1085–99 (1989).

41. Eyun, S. *et al.* Evolutionary History of Chemosensory-Related Gene Families across the Arthropoda. *Mol. Biol. Evol.* **34**, 1838–1862 (2017).

42. Croset, V. *et al.* Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. *PLoS Genet.* **6**, e1001064 (2010).

43. Robertson, H. M. The Insect Chemoreceptor Superfamily Is Ancient in Animals. *Chem. Senses* **40**, 609–614 (2015).

44. Lozano-Fernandez, J. *et al.* Increasing species sampling in chelicerate genomic-scale datasets provides support for monophyly of Acari and Arachnida. *Nat. Commun.* **10**, 2295 (2019).

45. Ache, B. W. & Young, J. M. Olfaction: Diverse Species, Conserved Principles. *Neuron* **48**, 417–430 (2005).

46. Hildebrand, J. G. & Shepherd, G. M. Mechanisms of olfactory discrimination: converging evidence for common principles across phyla. *Annu. Rev. Neurosci.* **20**, 595–631 (1997).

47. Sánchez-Gracia, A., Vieira, F. G. & Rozas, J. Molecular evolution of the major chemosensory gene families in insects. *Heredity* **103**, 208–216 (2009).

48. Vosshall, L. B. & Stocker, R. F. Molecular Architecture of Smell and Taste in *Drosophila*. *Annu. Rev. Neurosci.* **30**, 505–533 (2007).

49. Pelosi, P. Perireceptor events in olfaction. *J. Neurobiol.* **30**, 3–19 (1996).

50. Carr, A. L. & Roe, M. Acarine attractants: Chemoreception, bioassay, chemistry and control. *Pestic. Biochem. Physiol.* **131**, 60–79 (2016).

51. Sonenshine, D. & Roe, R. *Biology of ticks*. (Oxford University Press, 2013).

52. Foelix, R. F. Chemosensitive hairs in spiders. *J. Morphol.* **132**, 313–33 (1970).

53. Foelix, R. F. & Chu-Wang, I. W. The morphology of spider sensilla. II. Chemoreceptors. *Tissue Cell* **5**, 461–78 (1973).

54. Cerveira, A. M. & Jackson, R. R. Love is in the air: olfaction-based mate-odour identification by jumping spiders from the genus *Cyrba*. *J. Ethol.* **31**, 29–34 (2012).

55. Foelix, R. F., Rast, B. & Peattie, A. M. Silk secretion from tarantula feet revisited: alleged spigots are probably chemoreceptors. *J. Exp. Biol.* **215**, 1084–9 (2012).

56. Hallberg, E. & Hansson, B. S. Arthropod sensilla: Morphology and phylogenetic considerations. *Microsc. Res. Tech.* **47**, 428–439 (1999).

57. Kenning, M., Schendel, V., Müller, C. H. G. & Sombke, A. Comparative morphology of ultimate and walking legs in the centipede *Lithobius forficatus* (Myriapoda) with functional implications. *Zool. Lett.* **5**, 3 (2019).

58. Hallberg, E. & Skog, M. Chemosensory Sensilla in Crustaceans. in *Chemical Communication in Crustaceans* 103–121 (Springer, 2010).

59. Harzsch, S. & Krieger, J. Crustacean olfactory systems: A comparative review and a crustacean perspective on olfaction in insects. *Prog. Neurobiol.* **161**, 23–60 (2018).

60. Strausfeld, N. J. *Arthropod Brains: Evolution, Functional Elegance, and Historical Significance*. (Harvard University Press, 2012).

61. Buck, L. & Axel, R. A novel multigene family may encode odorant receptors: A molecular basis for odor recognition. *Cell* **65**, 175–187 (1991).

62. Bargmann, C. I. Comparative chemosensation from receptors to ecology. *Nature* **444**, 295–301 (2006).

63. Robertson, H. M., Warr, C. G. & Carlson, J. R. Molecular evolution of the insect chemoreceptor gene superfamily in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* **100 Suppl**, 14537–42 (2003).

64. Larsson, M. C. *et al.* Or83b Encodes a Broadly Expressed Odorant Receptor Essential for Drosophila Olfaction. *Neuron* **43**, 703–714 (2004).

65. Kwon, J. Y., Dahanukar, A., Weiss, L. A. & Carlson, J. R. The molecular basis of CO_2 reception in *Drosophila*. *Proc. Natl. Acad. Sci. USA* **104**, 3574–3578 (2007).

66. Robertson, H. M. Molecular Evolution of the Major Arthropod Chemoreceptor Gene Families. *Annu. Rev. Entomol.* **64**, 227–242 (2019).

67. Wicher, D. & Hansson, B. S. The *Drosophila* Carbon Dioxide Receptor as Key Regulator of Odor Valence. *Neuron* **97**, 996–997 (2018).

68. Freeman, E. G., Wisotsky, Z. & Dahanukar, A. Detection of sweet tastants by a conserved group of insect gustatory receptors. *Proc. Natl. Acad. Sci. USA* **111**, 1598–603 (2014).

69. Brand, P. *et al.* The origin of the odorant receptor gene family in insects. *Elife* **7**, (2018).

70. Missbach, C. *et al.* Evolution of insect olfactory receptors. *Elife* **3**, e02115 (2014).

71. Peñalva-Arana, D. C., Lynch, M. & Robertson, H. M. The chemoreceptor genes of the waterflea *Daphnia pulex*: many Grs but no Ors. *BMC Evol. Biol.* **9**, 79 (2009).

72. Chipman, A. D. *et al.* The first myriapod genome sequence reveals conservative arthropod gene content and genome organisation in the centipede *Strigamia maritima*. *PLoS Biol.* **12**, e1002005 (2014).

73. Hoy, M. A. *et al.* Genome Sequencing of the Phytoseiid Predatory Mite *Metaseiulus occidentalis* Reveals Completely Atomized Hox Genes and Superdynamic Intron Evolution. *Genome Biol. Evol.* **8**, 1762–1775 (2016).

74. Gulia-Nuss, M. *et al.* Genomic insights into the *Ixodes scapularis* tick vector of Lyme disease. *Nat. Commun.* **7**, 10507 (2016).

75. Saina, M. *et al.* A cnidarian homologue of an insect gustatory receptor functions in developmental body patterning. *Nat. Commun.* **6**, 6243 (2015).

76. Benton, R., Vannice, K. S., Gomez-Diaz, C. & Vosshall, L. B. Variant ionotropic glutamate receptors as chemosensory receptors in *Drosophila*. *Cell* **136**, 149–62 (2009).

77. Rimal, S. & Lee, Y. The multidimensional ionotropic receptors of *Drosophila melanogaster*. *Insect Mol. Biol.* **27**, 1–7 (2018).

78. Mayer, M. L. & Armstrong, N. Structure and Function of Glutamate Receptor Ion Channels. *Annu. Rev. Physiol.* **66**, 161–181 (2004).

79. Littleton, J. T. & Ganetzky, B. Ion channels and synaptic organization: analysis of the *Drosophila* genome. *Neuron* **26**, 35–43 (2000).

80. Tikhonov, D. B. & Magazanik, L. G. Origin and Molecular Evolution of Ionotropic Glutamate Receptors. *Neurosci. Behav. Physiol.* **39**, 763–773 (2009).

81. Stengl, M. Chemosensory Transduction in Arthropods. in *The Oxford Handbook of Invertebrate Neurobiology* 344–366 (Oxford University Press, 2019).

82. Rytz, R., Croset, V. & Benton, R. Ionotropic receptors (IRs): chemosensory ionotropic glutamate receptors in *Drosophila* and beyond. *Insect Biochem. Mol. Biol.* **43**, 888–97 (2013).

83. Abuin, L. *et al.* Functional architecture of olfactory ionotropic glutamate receptors. *Neuron* **69**, 44–60 (2011).

84. Ai, M. *et al.* Ionotropic glutamate receptors IR64a and IR8a form a functional odorant receptor complex in vivo in *Drosophila*. *J. Neurosci.* **33**, 10741–9 (2013).

85. Knecht, Z. A. *et al.* Distinct combinations of variant ionotropic glutamate receptors mediate thermosensation and hygrosensation in *Drosophila*. *Elife* **5**, (2016).

86. Enjin, A. *et al.* Humidity Sensing in *Drosophila*. *Curr. Biol.* **26**, 1352–8 (2016).

87. Joseph, R. M. & Carlson, J. R. *Drosophila* Chemoreceptors: A Molecular Interface Between the Chemical World and the Brain. *Trends Genet.* **31**, 683–695 (2015).

88. Stewart, S., Koh, T.-W., Ghosh, A. C. & Carlson, J. R. Candidate ionotropic taste receptors in the *Drosophila* larva. *Proc. Natl. Acad. Sci. USA* **112**, 4195–201 (2015).

89. Colbourne, J. K. *et al.* The ecoresponsive genome of *Daphnia pulex*. *Science* **331**, 555–61 (2011).

90. Stepanyan, R., Hollins, B., Brock, S. E. & McClintock, T. S. Primary Culture of Lobster (*Homarus americanus*) Olfactory Sensory Neurons. *Chem. Senses* **29**, 179–187 (2004).

91. Brockie, P. J., Madsen, D. M., Zheng, Y., Mellem, J. & Maricq, A. V. Differential expression of glutamate receptor subunits in the nervous system of *Caenorhabditis elegans* and their regulation by the homeodomain protein UNC-42. *J. Neurosci.* **21**, 1510–22 (2001).

92. Chen, Z., Wang, Q. & Wang, Z. The amiloride-sensitive epithelial Na+ channel PPK28 is essential for *Drosophila* gustatory water reception. *J. Neurosci.* **30**, 6247–52 (2010).

93. Lu, B., LaMora, A., Sun, Y., Welsh, M. J. & Ben-Shahar, Y. *ppk23*-Dependent Chemosensory Functions Contribute to Courtship Behavior in *Drosophila melanogaster*. *PLoS Genet.* **8**, e1002587 (2012).

94. Benton, R., Vannice, K. S. & Vosshall, L. B. An essential role for a CD36-related receptor in pheromone detection in *Drosophila*. *Nature* **450**, 289–93 (2007).

95. Gomez-Diaz, C. *et al.* A CD36 ectodomain mediates insect pheromone detection via a putative tunnelling mechanism. *Nat. Commun.* **7**, 11866 (2016).

96. Hanukoglu, I. & Hanukoglu, A. Epithelial sodium channel (ENaC) family: Phylogeny, structure-function, tissue distribution, and associated inherited diseases. *Gene* **579**, 95–132 (2016).

97. Ge, Y. & Elghetany, M. T. CD36: a multiligand molecule. *Lab. Hematol.* **11**, 31–7 (2005).

98. Vogt, R. G. & Riddiford, L. M. Pheromone binding and inactivation by moth antennae. *Nature* **293**, 161–3 (1981).

99. Flower, D. R., North, A. C. & Sansom, C. E. The lipocalin protein family: structural and sequence overview. *Biochim. Biophys. Acta* **1482**, 9–24 (2000).

100. Vieira, F. G. & Rozas, J. Comparative genomics of the odorant-binding and chemosensory protein gene families across the arthropoda: Origin and evolutionary history of the chemosensory system. *Genome Biol. Evol.* **3**, 476–490 (2011).

101. Leal, W. S., Nikonova, L. & Peng, G. Disulfide structure of the pheromone binding protein from the silkworm moth, *Bombyx mori*. *FEBS Lett.* **464**, 85–90 (1999).

102. Vieira, F. G., Sánchez-Gracia, A. & Rozas, J. Comparative genomic analysis of the odorant-binding protein family in 12 *Drosophila* genomes: purifying selection and birth-and-death evolution. *Genome Biol.* **8**, R235 (2007).

103. Gong, D.-P., Zhang, H.-J., Zhao, P., Xia, Q.-Y. & Xiang, Z.-H. The Odorant Binding Protein Gene Family from the Genome of Silkworm, *Bombyx mori*. *BMC Genomics* **10**, 332 (2009).

104. Jansen, S. *et al.* Structure of *Bombyx mori* chemosensory protein 1 in solution. *Arch. Insect Biochem. Physiol.* **66**, 135–145 (2007).

105. Angeli, S. *et al.* Purification, structural characterization, cloning and immunocytochemical localization of chemoreception proteins from *Schistocerca gregaria*. *Eur. J. Biochem.* **262**, 745–754 (1999).

106. Pelosi, P., Iovinella, I., Felicioli, A. & Dani, F. R. Soluble proteins of chemical communication: an overview across arthropods. *Front. Physiol.* **5**, 320 (2014).

107. Kitabayashi, A. N., Arai, T., Kubo, T. & Natori, S. Molecular cloning of cDNA for p10, a novel protein that increases in the regenerating legs of *Periplaneta americana* (American cockroach). *Insect Biochem. Mol. Biol.* **28**, 785–790 (1998).

108. Li, S. *et al.* Multiple functions of an odorant-binding protein in the mosquito *Aedes aegypti*. *Biochem. Biophys. Res. Commun.* **372**, 464–468 (2008).

109. Ishida, Y., Ishibashi, J. & Leal, W. S. Fatty acid solubilizer from the oral disk of the blowfly. *PLoS One* **8**, e51779 (2013).

110. Storch, J. & Xu, Z. Niemann-Pick C2 (NPC2) and intracellular cholesterol trafficking. *Biochim. Biophys. Acta* **1791**, 671–8 (2009).

111. Ishida, Y. *et al.* Niemann-Pick type C2 protein mediating chemical communication in the worker ant. *Proc. Natl. Acad. Sci. USA* **111**, 3847–52 (2014).

112. Nei, M. & Rooney, A. P. Concerted and Birth-and-Death Evolution of Multigene Families. *Annu. Rev. Genet.* **39**, 121–152 (2005).

113. Innan, H. & Kondrashov, F. The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* **11**, 97–108 (2010).

114. Conrad, B. & Antonarakis, S. E. Gene Duplication: A Drive for Phenotypic Diversity and Cause of Human Disease. *Annu. Rev. Genomics Hum. Genet.* **8**, 17–35 (2007).

115. Ward, P., Labandeira, C., Laurin, M. & Berner, R. A. Confirmation of Romer's Gap as a low oxygen interval constraining the timing of initial arthropod and vertebrate terrestrialization. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 16818–22 (2006).

116. Rota-Stabelli, O., Daley, A. C. & Pisani, D. Molecular timetrees reveal a Cambrian colonization of land and a new scenario for ecdysozoan evolution. *Curr. Biol.* **23**, 392–8 (2013).

117. Dunlop, J. A. & Selden, P. A. Calibrating the chelicerate clock: a paleontological reply to Jeyaprakash and Hoy. *Exp. Appl. Acarol.* **48**, 183–97 (2009).

118. Yarger, J. L., Cherry, B. R. & van der Vaart, A. Uncovering the structure–function relationship in spider silk. *Nat. Rev. Mater.* **3**, 18008 (2018).

119. Babb, P. L. *et al.* The *Nephila clavipes* genome highlights the diversity of spider silk genes and their complex expression. *Nat. Genet.* **49**, 895–903 (2017).

120. Sanggaard, K. W. *et al.* Spider genomes provide insight into composition and evolution of venom and silk. *Nat. Commun.* **5**, 3765 (2014).

121. Tobassum, S. *et al.* Nature and applications of scorpion venom: an overview. *Toxin Rev.* 1–12 (2018).

122. Grbić, M. *et al.* The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. *Nature* **479**, 487–92 (2011).

123. Ballesteros, J. A. & Sharma, P. P. A Critical Appraisal of the Placement of Xiphosura (Chelicerata) with Account of Known Sources of Phylogenetic Error. *Syst. Biol.* **syz011**, (2019).

124. Sharma, P. P. *et al.* Phylogenomic interrogation of arachnida reveals systemic conflicts in phylogenetic signal. *Mol. Biol. Evol.* **31**, 2963–84 (2014).

125. World Spider Catalog. *Nat. Hist. Museum Bern* online at http://wsc.nmbe.ch; version 20.0 (2019).

126. Macías-Hernández, N., López, S. de la C., Roca-Cusachs, M., Oromí, P. & Arnedo, M. A. A geographical distribution database of the genus *Dysdera* in the Canary Islands (Araneae, Dysderidae). *Zookeys* 11–23 (2016).

127. Arnedo, M. Radiation of the Spider Genus *Dysdera* (Araneae, Dysderidae) in the Canary Islands: Cladistic Assessment Based on Multiple Data Sets. *Cladistics* **17**, 313–353 (2001).

128. Arnedo, M. A., Oromí, P., Múrria, C., Macías-Hernández, N. & Ribera, C. The dark side of an island radiation: Systematics and evolution of troglobitic spiders of the genus *Dysdera* Latreille (Araneae:Dysderidae) in the Canary Islands. *Invertebr. Syst.* **21**, 623–660 (2007).

129. Pekár, S., Líznarová, E. & Řezáč, M. Suitability of woodlice prey for generalist and specialist spider predators: a comparative study. *Ecol. Entomol.* **41**, 123–130 (2016).

130. Řezáč, M. & Pekár, S. Evidence for woodlice-specialization in *Dysdera* spiders: behavioural versus developmental approaches. *Physiol. Entomol.* **32**, 367–371 (2007).

131. Řezáč, M., Pekár, S. & Lubin, Y. How oniscophagous spiders overcome woodlouse armour. *J. Zool.* **275**, 64–71 (2008).

132. Gorvett, H. Tegumental glands and terrestrial life in woodlice. *Proc. Zool. Soc. London* **126**, 291–314 (1956).

133. Sutton, S. L. *Woodlice*. (Pergamon Press, 1980).

134. Schmalfuss, H. Eco-morphological strategies in terrestrial isopods. *Symp. Zool. Soc. London* **53**, 49–63 (1984).

135. van Gestel, C. A. M., Loureiro, S. & Idar, P. Terrestrial isopods as model organisms in soil ecotoxicology: a review. *Zookeys* **801**, 127-162 (2018).

136. Hopkin, S. P., Martin, M. H. & Moss, S. J. Heavy metals in isopods from the supra-littoral zone on the Southern shore of the Severn Estuary, UK. *Environ. Pollut. Ser. B, Chem. Phys.* **9**, 239–254 (1985).

137. Singh, R., Gautam, N., Mishra, A. & Gupta, R. Heavy metals and living systems: An overview. *Indian J. Pharmacol.* **43**, 246 (2011).

138. Drobne, D. Terrestrial isopods-a good choice for toxicity testing of pollutants in the terrestrial environment. *Environ. Toxicol. Chem.* **16**, 1159–1164 (1997).

139. Dejean, A. Distribution of colonies and prey specialization in the ponerine ant genus *Leptogenys* (Hymenoptera: Formicidae). *Sociobiology* **29**, 293–299 (1997).

140. Hopkin, S. P. & Martin, M. H. Assimilation of zinc, cadmium, lead, copper, and iron by the spider *Dysdera crocata*, a predator of woodlice. *Bull. Environ. Contam. Toxicol.* **34**, 183–187 (1985).

141. Toft, S. & Macías-Hernández, N. Metabolic adaptations for isopod specialization in three species of *Dysdera* spiders from the Canary Islands. *Physiol. Entomol.* **42**, 191–198 (2017).

142. van den Bogaard, P. The origin of the Canary Island Seamount Province - New ages of old seamounts. *Sci. Rep.* **3**, 2107 (2013).

143. Kreitman, M. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature* **304**, 412–417 (1983).

144. Kreitman, M. E. & Aguadé, M. Excess polymorphism at the *Adh* locus in *Drosophila melanogaster*. *Genetics* **114**, 93–110 (1986).

145. Sackton, T. B. *et al.* Population Genomic Inferences from Sparse High-Throughput Sequencing of Two Populations of *Drosophila melanogaster*. *Genome Biol. Evol.* **1**, 449–465 (2009).

146. Langley, C. H. *et al.* Genomic Variation in Natural Populations of *Drosophila melanogaster*. *Genetics* **192**, 533–598 (2012).

147. Mackay, T. F. C. *et al.* The *Drosophila melanogaster* Genetic Reference Panel. *Nature* **482**, 173–178 (2012).

148. Soria-Carrasco, V. *et al.* Stick Insect Genomes Reveal Natural Selection's Role in Parallel Speciation. *Science* **344**, 738–742 (2014).

149. Stapley, J. *et al.* Adaptation genomics: the next generation. *Trends Ecol. Evol.* **25**, 705–712 (2010).

150. Ellegren, H. Comparative genomics and the study of evolution by natural selection. *Mol. Ecol.* **17**, 4586–4596 (2008).

151. Fumagalli, M. *et al.* Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure through Human Evolution. *PLoS Genet.* **7**, e1002355 (2011).

152. Sackton, T. B. *et al.* Dynamic evolution of the innate immune system in *Drosophila*. *Nat. Genet.* **39**, 1461–1468 (2007).

153. McTaggart, S. J., Obbard, D. J., Conlon, C. & Little, T. J. Immune genes undergo more adaptive evolution than non-immune system genes in *Daphnia pulex*. *BMC Evol. Biol.* **12**, 63 (2012).

154. Vuong, H. Q. & McFrederick, Q. S. Comparative Genomics of Wild Bee and Flower Isolated *Lactobacillus* Reveals Potential Adaptation to the Bee Host. *Genome Biol. Evol.* **11**, 2151–2161 (2019).

155. Faherty, S. L., Villanueva-Cañas, J. L., Blanco, M. B., Albà, M. M. & Yoder, A. D. Transcriptomics in the wild: Hibernation physiology in free-ranging dwarf lemurs. *Mol. Ecol.* **27**, 709–722 (2018).

156. Zhang, J., Chiodini, R., Badr, A. & Zhang, G. The impact of next-generation sequencing on genomics. *J. Genet. Genomics* **38**, 95–109 (2011).

157. Dominguez Del Angel, V. *et al.* Ten steps to get started in Genome Assembly and Annotation. *F1000Research* **7**, ELIXIR-148 (2018).

158. Yandell, M. & Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **13**, 329–342 (2012).

159. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).

160. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767–769 (2016).

161. Josek, T., Walden, K. K. O., Allan, B. F., Alleyne, M. & Robertson, H. M. A foreleg transcriptome for *Ixodes scapularis* ticks: Candidates for chemoreceptors and binding proteins that might be expressed in the sensory Haller's organ. *Ticks Tick. Borne. Dis.* **9**, 1317–1327 (2018).

162. Vizueta, J., Rozas, J. & Sánchez-Gracia, A. Comparative Genomics Reveals Thousands of Novel Chemosensory Genes and Massive Changes in Chemoreceptor Repertories across Chelicerates. *Genome Biol. Evol.* **10**, 1221–1236 (2018).

163. Clifton, B. D. *et al.* Rapid Functional and Sequence Differentiation of a Tandemly Repeated Species-Specific Multigene Family in *Drosophila*. *Mol. Biol. Evol.* **34**, 51–65 (2017).

164. Vizueta, J. *et al.* Evolution of chemosensory gene families in arthropods: Insight from the first inclusive comparative transcriptome analysis across spider appendages. *Genome Biol. Evol.* **9**, 178–196 (2017).

165. Altschul, S. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

166. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).

167. Simon, J.-C. *et al.* Genomics of adaptation to host-plants in herbivorous insects. *Brief. Funct. Genomics* **14**, 413–423 (2015).

168. Ngoc, P. C. T. *et al.* Complex Evolutionary Dynamics of Massively Expanded Chemosensory Receptor Families in an Extreme Generalist Chelicerate Herbivore. *Genome Biol. Evol.* **8**, 3323–3339 (2016).

169. Almeida, F. C., Sánchez-Gracia, A., Campos, J. L. & Rozas, J. Family size evolution in *Drosophila* chemosensory gene families: a comparative analysis with a critical appraisal of methods. *Genome Biol. Evol.* **6**, 1669–82 (2014).

170. Robertson, H. M., Baits, R. L., Walden, K. K. O., Wada-Katsumata, A. & Schal, C. Enormous expansion of the chemosensory gene repertoire in the omnivorous German cockroach *Blattella germanica*. *J. Exp. Zool. Part B Mol. Dev. Evol.* **330**, 265–278 (2018).

171. Qu, S.-X., Wang, X.-F., Li, H.-P., Luo, X. & Ma, L. A Gustatory Receptor Used for Rapid Detection of *Tyrophagus putrescentiae* in Fungi Hosts. *Sci. Rep.* **8**, 11425 (2018).

172. Bogusz, M. & Whelan, S. Phylogenetic Tree Estimation With and Without Alignment: New Distance Methods and Benchmarking. *Syst. Biol.* **66**, 218–231 (2017).

173. Schwager, E. E. *et al.* The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. *BMC Biol.* **15**, 62 (2017).

174. Suzuki, H. C. *et al.* Evolution of Gustatory Receptor Gene Family Provides Insights into Adaptation to Diverse Host Plants in Nymphalid Butterflies. *Genome Biol. Evol.* **10**, 1351–1362 (2018).

175. Iovinella, I. *et al.* Proteomic analysis of chemosensory organs in the honey bee parasite *Varroa destructor*: A comprehensive examination of the potential carriers for semiochemicals. *J. Proteomics* **181**, 131–141 (2018).

176. Renthal, R. *et al.* The chemosensory appendage proteome of *Amblyomma americanum* (Acari: Ixodidae) reveals putative odorant-binding and other chemoreception-related proteins. *Insect Sci.* **24**, 730–742 (2017).

177. Sánchez-Herrero, J. F. *et al.* The draft genome sequence of the spider *Dysdera silvatica* (Araneae, Dysderidae): A valuable resource for functional and evolutionary genomic studies in chelicerates. *Gigascience* **8**, giz099 (2019).

178. Mahler, D. L., Ingram, T., Revell, L. J. & Losos, J. B. Exceptional Convergence on the Macroevolutionary Landscape in Island Lizard Radiations. *Science* **341**, 292–295 (2013).

179. Storz, J. F. Causes of molecular convergence and parallelism in protein evolution. *Nat. Rev. Genet.* **17**, 239–250 (2016).

Anexo

A

Comparative analysis of tissue-specific transcriptomes in the funnel-web spider *Macrothele calpeiana* (Araneae, Hexathelidae)

PeerJ

Comparative analysis of tissue-specific transcriptomes in the funnel-web spider *Macrothele calpeiana* (Araneae, Hexathelidae)

Cristina Frías-López[1,2], Francisca C. Almeida[1,*], Sara Guirao-Rico[1,**], Joel Vizueta[1], Alejandro Sánchez-Gracia[1], Miquel A. Arnedo[2] and Julio Rozas[1]

[1] Departament de Genètica and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain
[2] Departament de Biologia Animal and Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain
[*] Current affiliation: Consejo Nacional de Investigaciones Científicas y Tecnológicas (CONICET), Departamento de Ecología, Genética y Evolución, Universidad de Buenos Aires, Intendente Güiraldes y Costanera Norte s/n, Pabellón II—Ciudad Universitaria, Capital Federal, Argentina
[**] Current affiliation: Centre for Research in Agricultural Genomics (CRAG) CSIC-IRTA-UAB-UB, Barcelona, Spain

ABSTRACT

The funnel-web spider *Macrothele calpeiana* is a charismatic Mygalomorph with a great interest in basic, applied and translational research. Nevertheless, current scarcity of genomic and transcriptomic data of this species clearly limits the research in this non-model organism. To overcome this limitation, we launched the first tissue-specific enriched RNA-seq analysis in this species using a subtractive hybridization approach, with two main objectives, to characterize the specific transcriptome of the putative chemosensory appendages (palps and first pair of legs), and to provide a new set of DNA markers for further phylogenetic studies. We have characterized the set of transcripts specifically expressed in putative chemosensory tissues of this species, much of them showing features shared by chemosensory system genes. Among specific candidates, we have identified some members of the iGluR and NPC2 families. Moreover, we have demonstrated the utility of these newly generated data as molecular markers by inferring the phylogenetic position *M. calpeina* in the phylogenetic tree of Mygalomorphs. Our results provide novel resources for researchers interested in spider molecular biology and systematics, which can help to expand our knowledge on the evolutionary processes underlying fundamental biological questions, as species invasion or biodiversity origin and maintenance.

Submitted 19 May 2015
Accepted 9 June 2015
Published 30 June 2015

Corresponding author
Julio Rozas, jrozas@ub.edu

Academic editor
Kimberly Bishop-Lilly

Additional Information and Declarations can be found on page 17

DOI **10.7717/peerj.1064**

Subjects Evolutionary Studies, Genetics, Genomics, Zoology
Keywords *De novo* transcriptome assembly, Molecular markers, Chemosensory system, RNA-seq, Mygalomorphae Phylogeny

INTRODUCTION

The funnel-web spider *Macrothele calpeiana* (family *Hexathelidae*) is a charismatic component of the European arthropod fauna. It belongs to the spider infraorder

How to cite this article Frías-López et al. (2015), Comparative analysis of tissue-specific transcriptomes in the funnel-web spider *Macrothele calpeiana* (Araneae, Hexathelidae). **PeerJ 3:e1064**; **DOI 10.7717/peerj.1064**

Mygalomorphae, which includes about 3,000 species of, among others, trap-door spiders, funnel-web spiders, and tarantulas (*Platnick, 2006*). *M. calpeiana* is a hairy, large spider that constructs extended and conspicuous funnel-web sheets close to the ground, and it is the only spider protected under European legislation (*Collins & Wells, 1987*). This spider is endemic to the southern Iberian Peninsula and was initially considered to be particularly vulnerable due to its close association with the highly threatened cork-oak forests found in the region (*Collins & Wells, 1987*). Subsequent studies, however, demonstrated that the species has a much wider distribution and could be frequently found in highly disturbed areas. In the last years, *M. calpeiana* has been introduced in European countries outside its natural range, probably associated with the commercial export of Spanish olive trees, raising some concerns about their possible impact on the invaded ecosystems (*Jiménez-Valverde, Decae & Arnedo, 2011*).

M. calpeiana is also an organism of particular interest in biogeographic studies. The *Macrothele* genus shows a highly disjointed distribution, with the bulk of its diversity in South-East Asia (21 species), a few species inhabiting tropical Africa (4 species) and only two known species in Europe, *M. calpeiana* itself and *M. cretica*, a Cretan endemic spider that is also of conservation concern. A recent phylogenetic study (*Opatova & Arnedo, 2014*) has revealed that the two European *Macrothele* species are not sister taxa, and that they most likely colonized independently Europe from Asia. Another interest in the genus relates to the venom toxins of some *Macrothele* spiders, which can be strong enough to cause envenomation, as in the case of some large Taiwanese *Macrothele* spiders (*Hung & Wang, 2004*). In fact, studies on the molecular structure and chemical properties of venom toxins (*Zeng, Xiao & Liang, 2003*; *Corzo et al., 2003*; *Satake et al., 2004*; *Yamaji et al., 2009*) have established the utility of *Macrothele* venom as cell growth inhibitors in cancer research (*Gao et al., 2005*; *Liu et al., 2012*).

The scarcity of genomic and transcriptomic data in chelicerates, which just cover a few species (*Grbić et al., 2011*; *Mattila et al., 2012*; *Cao et al., 2013*; *Clarke et al., 2014*; *Sanggaard et al., 2014*; *Posnien et al., 2014*) and the lack of tissue-specific transcript data in mygalomorphs, clearly limit the research on the molecular determinants of fundamental biological processes in this group of species. Within this context, with the aim of shedding light on the composition of Mygalomorph transcriptomes, we conducted the first RNA-seq study in one species of this group, *M. calpeiana*, including several tissues, and using a 454GS-FLX-based technology (*Prosdocimi et al., 2011*). The new sequence data will be an important, initial contribution to further basic, applied, and translational research in this non-model organism. Here we address two specific objectives: (i) to identify possible candidate chemosensory transcripts for future studies, and (ii) to provide new markers for further phylogenetic and evolutionary genomic-based studies in this group. As an example, we used some of the new generated transcripts to clarify the phylogenetic position of *M. calpeiana* in the Mygalomorph phylogeny.

The chemosensory system plays a key role in fundamental vital processes, including the localization of food, hosts, or predators and social communication; nevertheless, there are very few studies focused in non-insect species results (*Vieira & Rozas, 2011*;

Montagné et al., 2015), and almost unknown in mygalomorphs. In insects, the main molecular components of the chemosensory system are encoded by two main groups of gene families (*Sánchez-Gracia, Vieira & Rozas, 2009*; *Vieira & Rozas, 2011*; *Almeida et al., 2014*) the chemoreceptors and the secreted ligand-binding proteins. The first include the gustatory (GR), olfactory (OR), and ionotropic (IR) receptors, while the second group, known as ligand-binding families, are the odorant-binding protein (OBP), chemosensory protein (CSP), chemosensory type A and B (CheA/B), and probably some members of the Niemann-Pick disease type C2-related (NPC2) family (*Pelosi et al., 2014*). The preliminary analyses of the genomic sequences of the chelicerates *I. scapularis* (M Gulia-Nuss et al., 2015, unpublished data), *Stegodyphus mimosarum, Acanthoscurria geniculata*, (*Sanggaard et al., 2014*), *Mesobuthus martensii* (*Cao et al., 2013*), and *Tetranychus urticae* (*Grbić et al., 2011*), as well as in other arthropods, like the centipede *Strigamia maritima* (*Chipman et al., 2014*), revealed the absence of the typical insect OR and OBP gene families in these species.

Several experimental studies of chelicerates have identified the presence of specialised chemosensory hairs predominantly in the distal segment of the first pair of legs and in palps (*Foelix, 1970*; *Foelix & Chu-Wang, 1973*; *Kronestedt, 1979*; *Cerveira & Jackson, 2012*). In order to investigate the presence of transcripts related to the chemosensory system in spiders, we sequenced the specific transcriptomes of these two structures in *M. calpeiana*. To enrich our samples in tissue-specific transcripts, we built subtractive normalized cDNA libraries for each of these tissues separately. Additionally, for comparative purposes, we also analysed the ovary RNA-seq data. In this way, this study represents a starting-point to characterize the gene expression in the putative chelicerate chemosensory system structures.

Because of their low vagility and restricted distributions, mygalomorph spiders are well-suited for monitoring the ecological and evolutionary conservation status of terrestrial ecosystems (*Bond et al., 2006*), while at the same time are also highly threatened by habitat destruction (*Harvey, 2002*). To date, however, the lack of informative nuclear markers has limited research on these organisms and has hampered the assessment of their conservation or invasive species status. The method we employed here provides useful data for developing nuclear molecular markers to be used in other evolutionary genomic, phylogenetic, and phylogeographic studies of *Mygalomorphae*.

METHODS

Sample collection and preparation

Four adult females of the spider *Macrothele calpeiana* were collected (Junta de Andalucía, Spain; permission: SGYB-AFR-CMM) in two different localities in the southern Iberian Peninsula, namely Iznalloz (Granada, N37.36468 W3.47183, 1,011 m) (individuals MAC-GR1, MAC-GR2, MAC-GR3) and Finca de los Helechales, rd. Cabeza la Vaca (Huelva, N38.09032 W6.46621, 749 m) (individual CRBAMM000991). For each individual, palps, distal segments of the first pair of legs (denoted as legs), ovaries, brains and muscle tissues (from the rest of legs) were dissected and stabilized in RNA later (Applied Biosystems/Ambion).

Total RNA extraction and cDNA preparation

Each tissue was disrupted and homogenized separately using a rotor-stator homogenizer. Total RNA was extracted with the RNeasy midi kit (Qiagen, Hilden, Germany). For all dissected tissues, except the ovary, the protocol included a proteinase K digestion step in order to digest contaminant proteins. All samples were enriched in poly(A) mRNA prior to library preparation using the Oligotex RNA midi kit (Qiagen, Hilden, Germany).

The purified mRNA was used as a template for synthesizing the first cDNA strand using the SMARter PCR cDNA Synthesis Kit (Clontech, Mountain View, California, USA). In this protocol, a poly(A)-specific primer initiates the first strand synthesis of cDNA, thus selecting for polyadenylated RNA while simultaneously keeping the concentration of ribosomal RNA low. The resulting single stranded cDNA was amplified with the Advantage2 PCR kit (Clontech, Mountain View, California, USA), using 23 (brain, leg and muscle) and 20 (palp and ovary) amplification cycles. Double stranded cDNA was purified using CHROMA SPIN-1000 columns (Clontech, Mountain View, California, USA) and subsequently cleaved with *Rsa1* to generate shorter, blunt-ended cDNA fragments, which are necessary for adaptor ligation and subtraction. The digested cDNA were then purified using a standard phenol:chloroform:isoamyl extraction.

Subtractive hybridization and RNA sequencing

Transcripts expressed specifically in the palps, legs, and ovaries were enriched using the PCR-Select cDNA Subtraction Kit (Clontech, Mountain View, California, USA). This technique is based on a method of selective amplification of differently expressed sequences. We used leg, palp, and ovary cDNA as tester (samples of interest) and brain and muscle cDNAs samples as driver (transcripts exclusively for subtraction purposes) samples. According to the kit's protocol, the tester samples are subdivided into two aliquots that receive different adaptors. These aliquots are mixed to driver cDNA (in a higher concentration), denatured, and allowed to reanneal to form double chain cDNA. The process in repeated once, but with the two aliquots of tester cDNA mixed together and some more tester cDNA added. Then a PCR is done in a way that only double chain cDNA formed by fragments with different adaptors at each end will be amplified (i.e., cDNA formed by the hybridization of single chain cDNA from different tester aliquots). In this way, the sample is enriched with cDNA specific to the tester tissue since the tester cDNA that hybridizes with driver cDNA does not get amplified. The subtraction process also normalizes the library so that the frequencies of each unique cDNA became less unequal, increasing the chances of sequencing a large number of unique cDNAs. The subtracted cDNA products were treated with RNase (Qiagen, Hilden, Germany) and purified with QIAquick PCR Purification Kit (Qiagen, Hilden, Germany).

Two micrograms of subtracted cDNA from each tester tissue was prepared for sequencing on a 454/ Roche GS-FLX Titanium sequencer using three different MID tags, one for each tissue. Double-stranded cDNA was nebulized to generate 500-kb fragments and a shotgun library prepared for GS-FLX sequencing as per the manufacturer's instructions (Roche, Basel, Switzerland), which was run on a 1/4 picotitre plate region.

Read processing, handling, and *de novo* transcriptome assembly

We used *sffinfo* script (Roche's Newbler package; 454 SFF Tools) to extract the DNA sequences (FASTA format) and quality scores (FastQ format) independently for each MID tag from the SFF file. We removed adapters and putative contaminant sequences (upon the UniVec database and the *E. coli* genome sequence data) with SeqClean script (http://compbio.dfci.harvard.edu/tgi/software/), with parameters: -v <sequence of adapters> -c 8 -l 40 -x 95 -y 11 -M -L -s <database of contaminant sequences>. We trimmed low-quality bases at the ends of the reads and removed those shorter than 100bp or with a mean quality score (Q) below 20 using the NGS QC Toolkit (*Patel & Jain, 2012*).

First, we conducted a complete *de novo* assembly using all reads from the three tissues altogether in Newbler v2.6 GS (454 life Sciences, Roche Diagnostics) with parameters -urt -cDNA -Denovo -mol 100 -moi 95 -url. Subsequently, we used the contigs and the non-assembled reads (i.e., singletons) from this first step as input for a second assembly round in CAP3 (*Huang, 1999*), with parameters –o 60 –p 95. Redundant transcripts and putative isoforms were removed using cd-hit-est program, to generate a list of unique transcripts (*Fu et al., 2012*). We then used the gsMapper program (included in Newbler package) to map original (after filtering) reads (from the 3 tissues) to the unique transcripts, discarding all reads exhibiting hard clipping (more than 10% of read length) with an in-house Perl script.

Functional annotation

We carried out most of the functional annotation of the assembled transcripts with blast (v. 2.2.29) (*Altschul, 1997*; *Camacho et al., 2009*), Blast2GO (*Conesa et al., 2005*), InterProScan (*Jones et al., 2014*) and TRUFA (*Kornobis et al., 2015*). We first conducted a series of similarity-based searches with blastx (E-value cut-off 10^{-3}) against the NCBI non-redundant (NCBI-nr) database, retrieving the 5 hits with the lowest E-value for each query transcript. We then used Blast2GO and TRUFA to: (i) assign the Gene Ontology (GO) terms to each of these transcripts and determine the involved KEGG pathways (*Kanehisa & Goto, 2000*), (ii) identify particular protein domain structures in the sequenced transcripts using the InterProScan search engine, and (iii) determine which GO terms, InterPro domains, and KEGG pathways were significantly enriched in particular tissues by applying the Fisher's exact test and controlling by the False Discovery Rate (FDR) (*Benjamini & Hochberg, 1995*).

To determine the efficiency of the subtractive approach employed here to enrich samples with tissue specific transcripts, we estimated the fraction of assembled transcripts encoding for putative housekeeping (HK) genes (i.e., transcripts expected to be expressed across different tissues). For the analysis, we considered that a *M. calpeiana* transcript encodes a HK gene if we obtained a significant blastx hit (E-value cut-off 10^{-3}) against a database that includes all HK genes shared between humans (data set from *Eisenberg & Levanon, 2013*) and *Drosophila melanogaster* (data set from *Lam et al., 2012*) (which correspond to the 80% and 94% of the human and *Drosophila* HK genes, respectively; Table S1A).

Furthermore, we also estimated the number of transcripts that encode genes included in the CEG (Cluster of Essential Genes) database (a set of 458 Eukaryotic Orthologous Groups proteins identified by the Core Eukaryotic Genes Mapping Approach, CEGMA) (*Parra, Bradnam & Korf, 2007*; *Parra et al., 2009*). CEG proteins are highly conserved and present in a wide range of eukaryotic organisms, being therefore a good dataset to assess the reliability of our RNA sequencing and transcript annotation. *VennDiagram* R package was used to obtain all graphic representations of the logical relations (http://cran.r-project.org/web/packages/VennDiagram/index.html).

In order to identify putative *M. calpeiana* chemosensory related transcripts, we carried out an additional specific and customized search. We first built a protein database (CheDB) with vertebrate and insect sequences that match against the InterPro protein family signatures associated with chemosensory function (Table S1B). Then, we conducted a blastx search (E-value of 10^{-3}) using the assembled contigs as query against the CheDB database. To minimize the percentage of false positive results, we checked whether the candidate chemosensory transcripts from the blast searches truly encoded the Pfam HMM core profiles corresponding to chemosensory protein domains, using the programs HMMER (*Eddy, 2009*) (E-value of 10^{-3}) and InterProScan. Only *M. calpeiana* transcripts with positive hits in this second search step were unequivocally annotated as putative chemosensory genes. Finally, we also ran an additional tblastn search (E-value of 10^{-3}) of a set of proteins annotated as chemosensory in currently available chelicerate genomes—the common house spider *Parasteatoda tepidariorum* (https://www.hgsc.bcm.edu/arthropods/common-house-spider-genome-project), the social spider *Stegodyphus mimosarum* (*Sanggaard et al., 2014*), the mygalomorph spider *Acanthoscurria geniculate* (*Sanggaard et al., 2014*), the scorpion *Mesobuthus martensii* (*Cao et al., 2013*), and the tick *Ixodes scapularis* (https://www.vectorbase.org/) against *M. calpeiana* transcripts. In this last search, we also included as queries the translated sequences of the transcripts already identified as candidate *M. calpeiana* chemosensory genes in the first searches. In order to exclude spurious homologs caused by short-length false-positive hits, we only considered for further analyses those transcripts whose blast alignments span either at least 2/3 of the total number of amino acids of the query proteins or those covering at least 80% of the transcript length.

Phylogenetic analysis

To determine the utility of the newly sequenced transcripts as markers for molecular phylogenetics, we applied them to study the phylogenetic position of *M. calpeiana* in the tree of Mygalomorphs, a currently unresolved question. As a starting point, we used the phylogenetic analysis reported in *Bond et al. (2014)*. In particular, we first retrieved the amino acid data of all 16 mygalomorph and 3 non-mygalomorph outgroup species (*Stegodyphus*, *Hypochilus* and *Liphistius*) from the matrix d327 (44 taxa; 327 genes; 110,808 amino acid positions). Then, we searched for putative homologs of these 327 genes in *M. calpeiana* transcripts using the blastp program. For this analysis, we obtained the conceptual translation of the transcript sequences (in all six frames) using TransDecoder (version r20140704) as implemented in the Trinity software (*Haas et al., 2013*). We selected

all *Macrothele* translated amino acid sequences that produced a positive blast hit with an E-value $< 10^{-15}$ and with local alignment length >80 amino acids (i.e., in order to maximize the probability of using 1:1 orthologues). Then, we aligned each of these selected translated sequences of *M. calpeiana* with their corresponding homologs in the 19 chosen species (a single multiple sequence alignment, MSA, per gene) using MAFFT (option–merge) (*Katoh & Standley, 2013*). Finally, we concatenated all individual MSA with amino acid data in at least 50% of the species.

We also built family specific MSA with amino acid sequences of NMDA-ionotropic glutamate receptors (NMDA-iGluR) and with members of the Niemann-Pick C disease 2 (NPC2) family, to investigate the phylogenetic relationships between the candidate *M. calpeiana* transcripts and some representatives of these two families in arthropods. We included in these MSA the proteins already annotated in *D. melanogaster* (hexapod), *S. maritima* (myriapod) and *I. scapularis* (chelicerate), as well as the NPC2 genes expressed in *Apis melifera* and *Camponotus japonicus* antenna (*Pelosi et al., 2014*). For iGluR (including IR8a/IR25a proteins) we prepared two different MSA, one for each functional domain. We used HMMER and the Pfam profiles of these two domains (PF01094 "ANF_receptor," and PF00060 "Lig_chan") to identify and trim separately the extracellular amino-terminal and the ligand-gated ion channel domains, which were used to build two separate MSA (and separate trees) with HMMERALIGN.

We conducted all phylogenetic reconstructions by maximum likelihood (ML) using the PROTGAMMAWAG model in the program RAxML version 8 (*Stamatakis, 2014*). We carried out a multiple non-parametric bootstrap analysis (500 bootstrap runs) to obtain node support values.

RESULTS AND DISCUSSION

RNA-seq of *Macrothele calpeiana*

We sequenced a total of 164,111 raw reads across the three tester samples (i.e., leg, palp, and ovary), with a N50 value of 409bp (Table 1). After trimming, cleaning and removing very short reads (less than 100bp), we obtained a final set of 128,816 reads, which was used for further analyses. Our two-step *de novo* assembly strategy (applying Newbler v 2.6, and subsequently CAP3) yielded a total of 3,705 contigs (N50 of 647bp), composed by more than one read, plus 3,560 singletons. After running the cd-hit-est and gsMapper software these contigs clustered into 6,696 unique sequences (i.e., putative *M. calpeiana* individual coding genes), of which 3,467 corresponds to contigs assembled by more than one read (i.e., excluding singletons) (Table 1; Table S2). Table 2 and Table S3 show the distribution of these 6,696 (and also the 3,467) unique sequences across tissues. *M. calpeiana* reads data are available at the Sequence Read Archive (SRA) database under the accession numbers SRA: SRS951615, SRA: SRS951616 and SRA: SRS951618 (Bioproject number: PRJNA285862).

RNA-seq quality and functional annotation

We investigated the quality of our tissue specific transcriptome by a series of similarity-based searches of our transcripts against sequences in the NCBI-nr database. As expected,

Table 1 Summary of RNA-seq data and assembly.

Raw number of reads	164,111
N50	409
Reads used in the Newbler assembly[a]	128,818
Assembled reads	122,183
Isotigs (number of singletons)	3,635 (6,614)
N50 (Isotigs)	601
CAP3 assembly	
Contigs (number of singletons >100nuc)	3,705 (3,560)
N50 (Contigs)	647
Unique sequences[b]	
Total number of sequences (transcripts)	6,696
N50	455
Coverage	14.33X
Reads mapped	95,250
Sequences (excluding singletons)	3,467
N50	613
Coverage	22.94X
Reads mapped	90,267

Notes.
[a] Number of reads after trimming, cleaning and excluding short reads.
[b] Number of reads after clustering and mapping filtering.

Table 2 Summary of RNA-seq data and assembly per tissue.

	Leg	Palp	Ovary	Total
Driver[a]	Muscle	Muscle	Brain	
Raw number of reads	59,232	54,321	50,558	164,111
N50	404	405	419	409
Reads used in assembly[b]	46,474	41,545	40,799	128,818
N50	362	364	378	368
Unique sequences (transcripts)[c]	2,705	3,798	1,796	6,696
Longest transcript (in nucleotides)	3,053	3,057	4,116	4,116
HK, housekeeping sequences	426	638	328	1,005
CEG sequences	385	547	236	789
Sequences excluding HK-CEG genes	2,139	2,952	1,369	5,390
Sequences with GO annotation	1,147	1,612	816	2,619
Sequences within Interpro	1,464	1,966	988	3,353
Sequences within KEGG	389	509	173	776
Sequences with functional annotation[d]	1,704	2,363	1,152	3,970
Sequences with annotation[e]	2,060	2,915	1,428	4,978

Notes.
[a] Driver of subtractive cDNA library.
[b] Number of reads after trimming, cleaning and excluding short reads.
[c] Considering the total ($n = 6,696$) data set.
[d] GO, Interpro or KEGG hits.
[e] GO, Interpro, KEGG or blast hits.

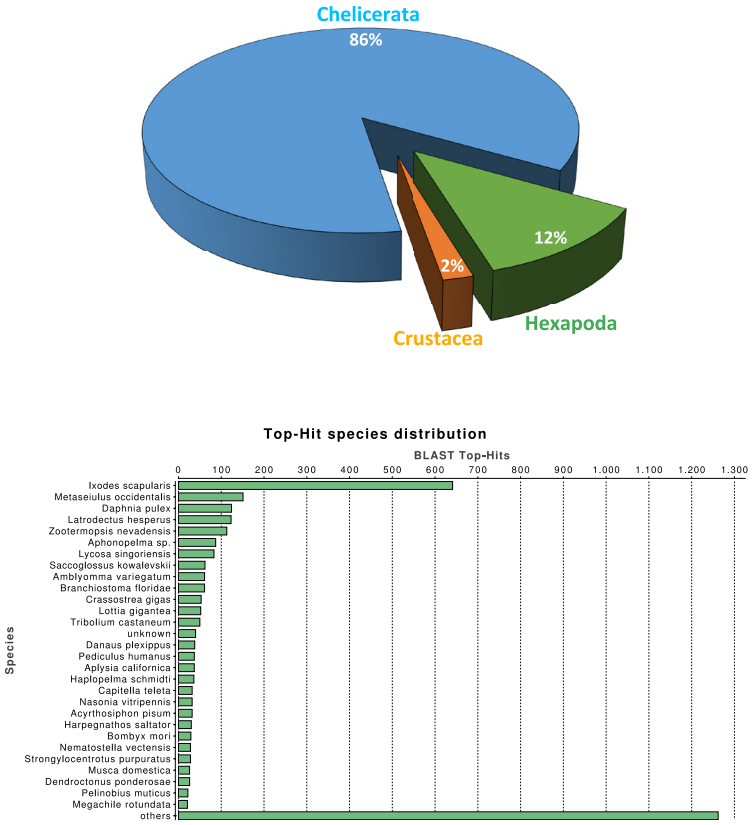


Figure 1 Macrothele taxonomic distribution. Taxonomic distribution of the 6,696 transcripts with significant blast hits against the NCBI-nr data base (using the top-hit; cut-off E-value of 10^{-3}) by means of the Blast2GO package (4,399 transcripts with blast hit). (A) Distribution of the top-hits across arthropod groups (29.4% of the transcripts with blast hit). (B) Top-hit species distribution.

the single largest category of top blast hits (blastx E-value cut-off 10^{-3}), corresponding to 25.3% of top blast hits, was to chelicerate protein coding genes, followed by hits to other arthropod species (4.1%). Within the Arthropoda, hits within Hexapoda represents about 12% (Fig. 1A), while *Ixodes scapularis* is the species receiving the majority of hits (Fig. 1B).

Overall, 2,619, 3,353 and 776 out of the 6,696 identified transcripts have a GO, InterPro, or KEGG associated term, respectively (Table 2); in total 4,978 of them (74.3%) have some functional annotation information. We analysed the distribution of GO terms (at GO level 2) across the 2,619 *M. calpeiana* transcripts sequences with GO annotation. We found that

the most frequent GO terms present in this sample are "metabolic" and "cellular processes" within the biological process domain (BP), and "binding" and "catalytic activities" within molecular function domain (MF). The distribution of GO terms in the complete data set (2,619 GO terms; Fig. 2) and in the data set excluding singleton sequences (1,734 GO terms; Fig. S1) is not significantly different (two tailed FET, P-value $= 0.592$ and 0.757 for BP and MF, respectively). Hence, we used the complete dataset for further functional annotation analyses.

Tissue-specific expression

With our subtractive approach we aimed to enrich a number of tissue-specific transcripts. We detected 1,005 transcripts annotated as housekeeping genes (Table 2) and 789 transcripts with putative homology to 290 of 458 CEG members of the CEGs dataset. Out of the 789 transcripts with CEG homologs, 488 are also annotated as HK genes (Fig. S2 and Tables S3–S5). Despite the finding of about 15% of HK and CEG genes, the largest proportion of them are located at the intersection of the Venn diagram (Figs. 3C and 3D), indicating that tissue-specific transcripts should reliably represent tissue-specific functions. After excluding these likely ubiquitously expressed genes, the remaining sample ($n = 5,390$ transcripts; 1,523 with GO annotation) exhibits the desired tissue-specific expression profile. In fact, the distributions of GO terms including (2,619 transcripts) or not (1,523 transcripts) HK/CEG genes are significantly different from each other (two tailed P-value < 0.018 for the most frequent GO categories within BP and MP) (Fig. 2).

To gain further insight into transcript function, we compared transcript expression across legs, palps, and ovaries (Fig. 3; Fig. S3). We found a high proportion of transcripts shared between leg and palp (1,112 and 848, including or not HK and CEG genes, respectively), and a few between these tissues and ovary (Figs. 3A and 3B). This result was expected given the ontogenetic similarities of legs and palps.

The overrepresentation analysis of the GO terms across the different Venn diagram sections (Table S3; see also Fig. 3E) detected 26 significant overrepresented GO terms in legs-palps (sections I, II and IV) or ovary transcripts (sections III, V, VI and VII) after the FDR (Fig. 4; Table S6A and Fig. S4). For instance, the GO terms "cation binding," "metal ion binding," and "oxidation–reduction process" are clearly overrepresented in legs-palps specific transcripts (P-value $< 6.9 \times 10^{-8}$). These significant differences are also found in comparisons involving only section III (i.e., considering only ovary-specific transcripts instead of all ovary-transcripts), or only section IV (considering only specific transcripts shared between leg and palp) (results not shown). Indeed, the major over- or underrepresentation effect appears in individual sections III and IV (Table S6).

To investigate the biological pathways that are differently expressed among the studied tissues, we analysed the distribution of transcripts associated with different KEGG terms (Tables S3 and S7). Again, we found significant differences between transcripts expressed exclusively in legs and/or palps (sections I, II, and IV) and the ovary-expressed transcripts (sections III, V, VI, and VII) (two tailed FET, P-value of 2.6×10^{-3}). For instance, we detected 3 KEGG pathways (Tropane, piperidine and pyridine alkaloid biosynthesis;

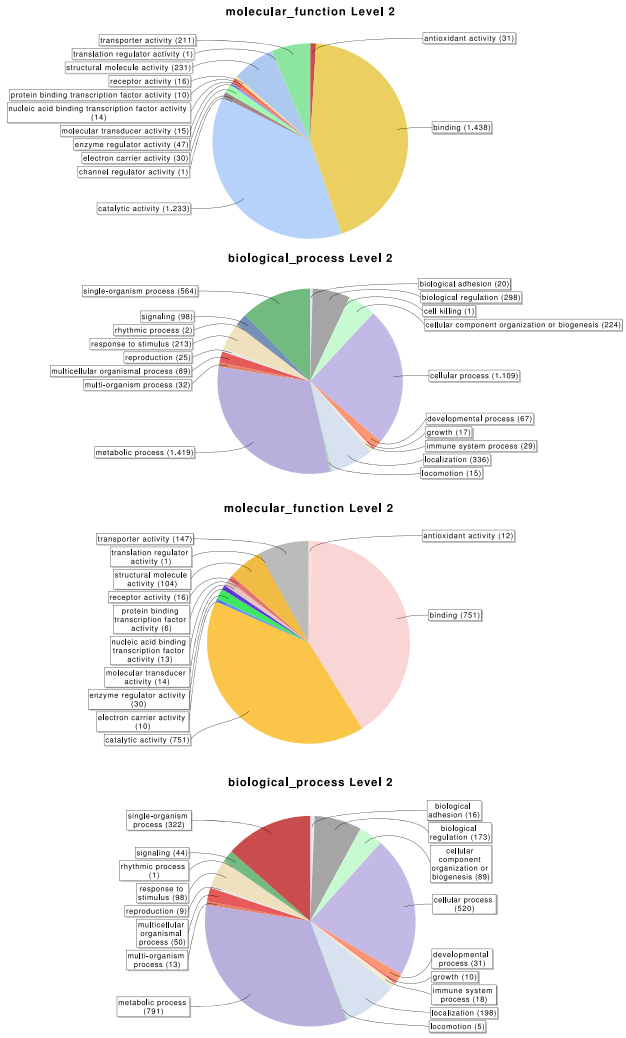


Figure 2 Distribution of the Gene Ontology (GO) terms associated with the complete set of *M. calpeiana* transcripts (2,619 transcripts with GO annotations over 6,696 sequences). (A) MF, molecular function. (B) BP, Biological process. Distribution GO terms excluding transcripts encoding HK or CEG genes (1,523 transcripts with GO annotations over 5,390 sequences). (C) MF, molecular function. (D) BP, Biological process.

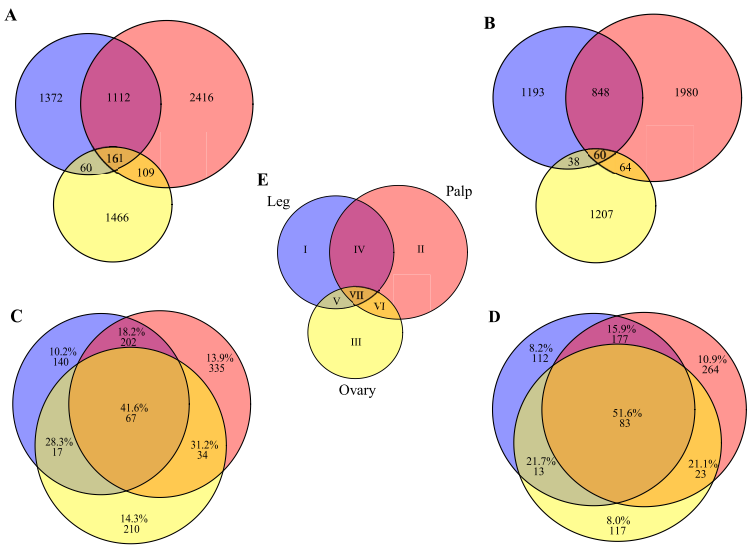


Figure 3 Transcript distribution across tissues. Venn diagrams showing the number of sequences expressed specifically in each tissue or in their intersections (blue, ochre and yellow indicate leg, palp and ovary, respectively). (A) All transcripts ($n = 6,696$). (B) Transcripts excluding putative housekeeping or CEG genes ($n = 5,390$). (C) Number and percentage of transcripts encoded by housekeeping genes ($n = 1,005$). (D) Number and percentage of transcripts with homologs included in the CEG database ($n = 789$). The area of each Venn diagram section is approximately proportional to the number of transcripts (A and B), or to the particular fraction value (C and D). (E) Roman numerals used to designate the different sections.

Tryptophan metabolism; and Tyrosine metabolism) specifically expressed in sections I, II and IV; none of the 11 detected transcripts of these three pathways had ovary expression (Table S7). Actually, these pathways are not directly related to chemosensory function. It has been shown that the golden orb web spider *Nephila antipodiana* (Walckenaer) coats its web with an alkaloid (2-pyrrolidinone), which apparently provides protection against ant invasion (*Zhang et al., 2012*). *Macrothele* large funnel-webs are equally exposed to predators, both insects and small vertebrates, and hence the use of a chemical defense against invaders would be highly advantageous. Further studies on the presence of these chemical clues on the funnel-webs are needed to confirm this hypothesis.

Chemosensory-related genes

As a starting point for the identification of chemosensory organs in *M. calpeiana*, we studied two features commonly present in the chemosensory-related proteins, the existence of a signal peptide (characteristic of soluble binding proteins such as insect and vertebrate OBP, and the NPC2, CSP, and CheA/B), and the presence of a transmembrane domain (characteristic of all chemosensory receptors, such as insect and vertebrate ORs,

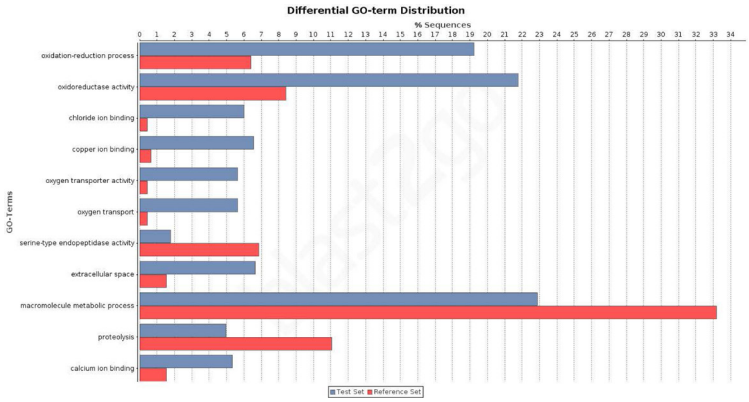


Figure 4 Differential distribution of GO terms across tissues. Differential distribution of the GO terms of the transcripts from leg and palp (Venn sections I, II and IV; in blue) and ovary (sections III, V, VI and VII; in red). Analysis conducted excluding HK and CEG encoding genes (1,523 transcripts over 5,390).

GRs and IRs). For that, we searched for a putative tissue-specific overrepresentation of such features in legs and palps (the candidate chemosensory structures in spiders) among the 3,353 transcripts with InterPro annotation. We found a significant over-representation of the signal peptide-encoding transcripts in legs-palps specific transcripts (Venn sections I, II and IV against the rest) (two tailed FET, P-value of 6.9×10^{-3}), being especially evident for transcripts shared between palps and legs tissues (Venn section IV; two tailed FET, P-value of 9.7×10^{-7}). Remarkably, the percentage of transcripts with signal peptide in section IV of the Venn diagram (transcripts expressed in both legs and palps, but not in ovary) is 27.8% (Fig. 5A), while the 40.6% of leg-specific transcripts have at least one transmembrane domain (Fig. 5B). Given that these features are not completely exclusive of chemosensory genes it is difficult to clearly assess whether these differences may reflect true differences in the chemosensory role of these tissues (see also Fig. S5).

The specific blast searches for chemosensory genes against the CheDB database detected several candidate transcripts. Nevertheless, the examination of the conceptual translation of these transcripts using HMM profiles showed that only seven candidates (two IR and five NPC2; Table S3) have the specific molecular signature of a chemosensory protein domain. Almost all the other candidates either exhibit non-chemosensory domain signatures or yielded no significant results in the search against HMM profiles. The two putative IR transcripts are specifically expressed in palps and each of them encodes a different Pfam domain characteristic of these receptors (*Croset et al., 2010*), the extracellular amino-terminal domain (PF01094; transcript Mcal_4794) and the ligand-gated ion channel domain (PF00060; transcript Mcal_5646). The closest related proteins of the *M. calpeiana* transcripts in the CheDB database correspond with two *S. mimosarum* predicted proteins annotated as "Glutamate receptor, ionotropic kainate 2" products (GenBank accessions

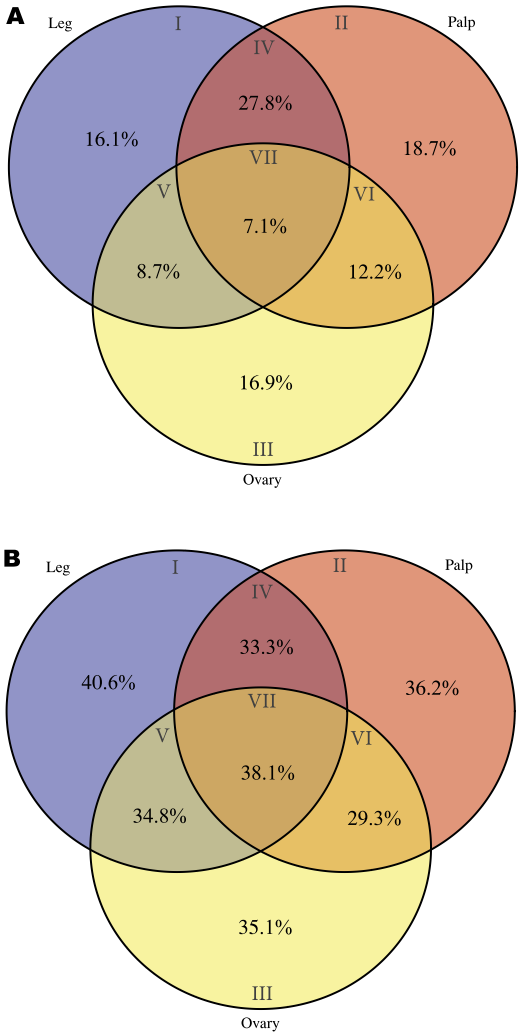


Figure 5 Distribution of specific interpro domains across tissues. Venn diagrams showing the percentage of specific interpro domains across tissues (the different Venn sections are indicated in roman numbers). Analysis conducted excluding HK and CEG encoding genes (2,364 transcripts with Interpro annotation over 5,390). (A) Signal peptide domain. (B) Transmembrane domain.

KFM81344 and KFM59881, 48% and 67% of identity, with Mcal_4794 and Mcal_5646, respectively). Nevertheless, we cannot rule out that the two *M. calpeiana* transcripts were in fact two fragments of the same iGluR gene since KFM59881 is also a partial product that only includes the "Lig_chan" domain. Besides, the rest of best-hits in blast searches using these two *M. calpeiana* transcripts as queries correspond to kainate (KA) receptors followed by α-amino-3-hydroxy-5-methyl-4-isoxazole propionate (AMPA) members in other arthropod species. The phylogenetic trees of the members of these subfamilies in arthropods (built separately for each protein domain; see 'Methods') show that the translated proteins of Mcal_4794 and Mcal_5646 group in the same clade with some KA receptors of insects, centipedes or ticks (Figs. S6A and S6B), again suggesting their putative role in synaptic transmission and regulation (i.e., it would not be a chemosensory receptor).

The products of three of the five putative NPC2 encoding transcripts constitute a *M. calpeiana* specific monophyletic clade in the NPC2 family tree (Fig. S6C) and are specifically expressed in ovary, which is suggestive of a non-chemosensory function. The other two NPC2 are expressed in palp and legs (Mcal_1484) or palp-specific (Mcal_6333). Both encoding proteins are relatively distant to the *Apis mellifera* and *Camponotus japonicus* antennal expressed NPC2, being more related to some *I. scapularis* and *S. maritima* members as well as with the ovarian clade of NPC2. In light of these results, the possible chemosensory function of these proteins in palps and legs remains to be elucidated. These results strongly encourage further functional analyses to determine the putative chemosensory role of these NPC2 genes specifically expressed in palps and legs.

Recent genome sequencing projects have revealed that chelicerate genomes contain numerous copies of ionotropic (IR) and insect-like gustatory (GR) receptors, which are the principal candidates to perform chemoreceptor functions in these species. The apparent absence of genes belonging to these families specifically expressed in *M. calpeiana* palp/leg tissues might be explained by low sequence coverage. Many of these receptors are probably encoded by low expressed genes, and their detection might need more extensive sequencing. However, to date, there is no other study of the specific expression of either these receptors or other chemosensory family members in different tissues of a chelicerate. Given the life-style of *M. calpeiana*, i.e., it builds funnel-shaped webs, which it uses to trap prey, we cannot rule out the possibility of a residual role of a chemoreceptor system in favour of mechanoreception in this species. New deep sequencing transcriptomic data from other spider species are needed to answer this question. In fact, our preliminary results from tissue specific transcriptomes in *Dysdera silvatica* (Araneae, Haplogynae) (J Vizueta et al., 2015, unpublished data) indicate that members IRs and GRs families are specifically expressed in leg and palp tissues, suggesting their putative role in chemoreception in nocturnal running hunter spiders.

Mygalomorph phylogeny

From the data matrix d327 of *Bond et al. (2014)*, we built a new MSA with information of *M. calpeiana* obtained from our transcriptome analysis. We have filtered the data in order to include high quality homologous data with high coverage per taxon. Our final MSA

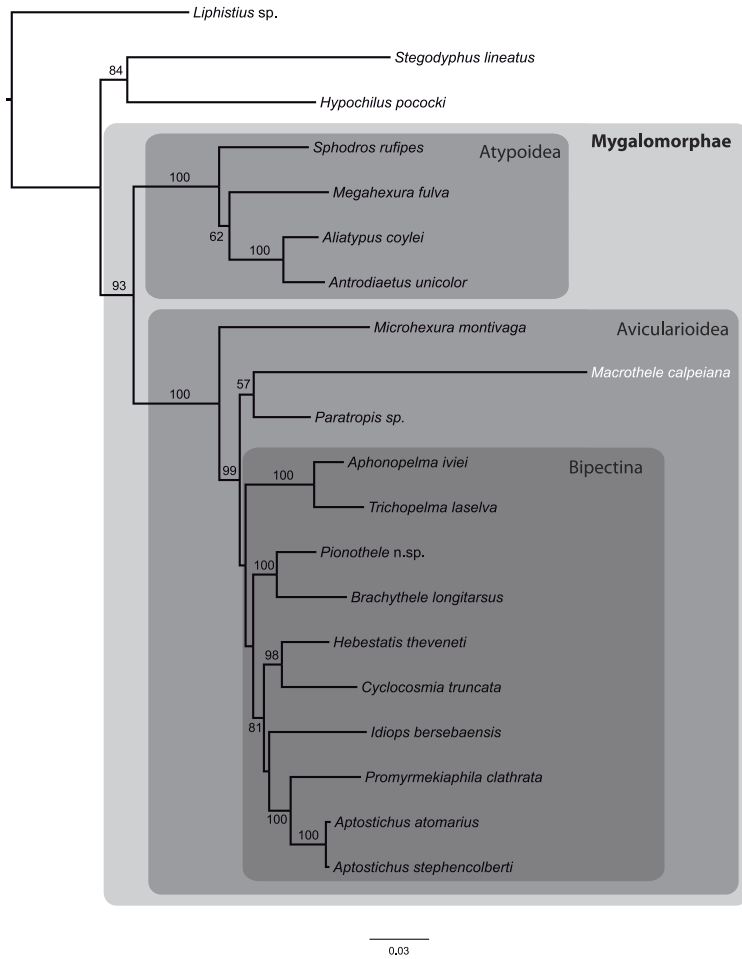


Figure 6 Phylogenetic relationships of major Mygalomorphae lineages sampled. ML tree showing the phylogenetic relationships of major Mygalomorphae lineages sampled. The analysis is based on a supermatrix of 35 putative orthologs (4,531 amino acids). Numbers indicate bootstrap support values >50%.

comprises 17 Mygalomorph species (including *M. calpeiana*) and 3 non-mygalomorph outgroups (20 taxa; 35 genes; 4,531 amino acids; Table S8), with an average taxa coverage of 17.1. Our ML phylogenetic tree, rooted using *Liphistus* as an outgroup, mirrors those reported in *Bond et al. (2014)* and shows *M. calpeiana* as the sister lineage of the genus *Paratropis* (Fig. 6), albeit with low node support (57%), as part of the non-Bipectina

Avicularioidea. Interestingly, in a recent study focused on the phylogenetic relationship and biogeographic origins of the genus *Macrothele* (*Opatova & Arnedo, 2014*) based on a denser taxonomic sampling but lower gene coverage (3 genes), a similar position of *Macrothele*, within the Avicularioidea but outside the Bipectina lineage, was also recovered.

CONCLUSIONS

The tissue specific transcriptome presented here provides a novel resource for *Macrothele* researchers, and for people interested in spider systematics and molecular biology. Having ovary and non-ovary expressed transcripts-based markers, which may potentially differ in their evolutionary rates, can become instrumental for further studies aiming to understand the evolutionary processes acting at different time-scales, such as biological invasions, secondary gene flow or speciation, and to implement successful conservation polices; in particular, we have demonstrated the utility of these newly generated data by inferring the phylogenetic position of *M. calpeiana* in the Mygalomorphae tree. Moreover, our tissue-specific gene expression study represents a starting point to understanding the chemosensory system in spiders and, in general, in chelicerates.

ACKNOWLEDGEMENT

We thank Centres Científics i Tecnològics de la Universitat de Barcelona for the sequencing facilities.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Grants from the Ministerio de Educación y Ciencia of Spain (BFU2010-15484 and CGL2013-45211 to JR, and CGL2012-36863 to MAA), and from the Comissió Interdepartamental de Recerca i Innovació Tecnològica of Spain (2009SGR-1287; 2014SGR1055; 2014SGR1604). JR and MAA were partially supported by ICREA Academia (Generalitat de Catalunya). CF-L was supported by an IRBio fellowship (Universitat de Barcelona), FCA by a Juan de la Cierva postdoctoral fellowship (Spanish Ministerio de Economía y Competitividad; JCI-2008-3456), and SG-R and AS-G by a grant under the program Beatriu de Pinós (Generalitat de Catalunya, 2010BP-A 00438 and 2010BP-B 00175, respectively). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Ministerio de Educación y Ciencia of Spain: BFU2010-15484, CGL2013-45211, CGL2012-36863.
Comissió Interdepartamental de Recerca i Innovació Tecnològica of Spain: 2009SGR-1287, 2014SGR1055, 2014SGR1604.
ICREA Academia (Generalitat de Catalunya).
IRBio fellowship (Universitat de Barcelona).

Juan de la Cierva postdoctoral fellowship (Spanish Ministerio de Economía y Competitividad): JCI-2008-3456.
Beatriu de Pinós postdoctoral fellowships (Generalitat de Catalunya): 2010BP-A 00438, 2010BP-B 00175.

Competing Interests
Julio Rozas is an Academic Editor for PeerJ.

Author Contributions
- Cristina Frías-López performed the experiments, analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Francisca C. Almeida and Sara Guirao-Rico performed the experiments, analyzed the data, reviewed drafts of the paper.
- Joel Vizueta analyzed the data, prepared figures and/or tables, reviewed drafts of the paper.
- Alejandro Sánchez-Gracia analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Miquel A. Arnedo conceived and designed the experiments, analyzed the data, contributed reagents/materials/analysis tools, reviewed drafts of the paper.
- Julio Rozas conceived and designed the experiments, analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.

Field Study Permissions
The following information was supplied relating to field study approvals (i.e., approving body and any reference numbers):

Field permission from the Junta de Andalucia (Spain); reference: SGYB-AFR-CMM.

DNA Deposition
The following information was supplied regarding the deposition of DNA sequences:
http://www.ncbi.nlm.nih.gov/bioproject/PRJNA285862.

Supplemental Information
Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.1064#supplemental-information.

REFERENCES
Almeida FC, Sánchez-Gracia A, Campos JL, Rozas J. 2014. Family size evolution in *Drosophila* chemosensory gene families: a comparative analysis with a critical appraisal of methods. *Genome Biology and Evolution* **6**:1669–1682 DOI 10.1093/gbe/evu130.

Altschul S. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**:3389–3402 DOI 10.1093/nar/25.17.3389.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**:289–300 DOI 10.2307/2346101.

Bond JE, Beamer DA, Lamb T, Hedin M. 2006. Combining genetic and geospatial analyses to infer population extinction in mygalomorph spiders endemic to the Los Angeles region. *Animal Conservation* **9**:145–157 DOI 10.1111/j.1469-1795.2006.00024.x.

Bond JE, Garrison NL, Hamilton CA, Godwin RL, Hedin M, Agnarsson I. 2014. Phylogenomics resolves a spider backbone phylogeny and rejects a prevailing paradigm for orb web evolution. *Current Biology: CB* **24**:1765–1771 DOI 10.1016/j.cub.2014.06.034.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**:421 DOI 10.1186/1471-2105-10-421.

Cao Z, Yu Y, Wu Y, Hao P, Di Z, He Y, Chen Z, Yang W, Shen Z, He X, Sheng J, Xu X, Pan B, Feng J, Yang X, Hong W, Zhao W, Li Z, Huang K, Li T, Kong Y, Liu H, Jiang D, Zhang B, Hu J, Hu Y, Wang B, Dai J, Yuan B, Feng Y, Huang W, Xing X, Zhao G, Li X, Li Y, Li W. 2013. The genome of *Mesobuthus martensii* reveals a unique adaptation model of arthropods. *Nature Communications* **4**:Article 2602 DOI 10.1038/ncomms3602.

Cerveira AM, Jackson RR. 2012. Love is in the air: olfaction-based mate-odour identification by jumping spiders from the genus Cyrba. *Journal of Ethology* **31**:29–34 DOI 10.1007/s10164-012-0345-x.

Chipman AD, Ferrier DEK, Brena C, Qu J, Hughes DST, Schröder R, Torres-Oliva M, Znassi N, Jiang H, Almeida FC, Alonso CR, Apostolou Z, Aqrawi P, Arthur W, Barna JCJ, Blankenburg KP, Brites D, Capella-Gutiérrez S, Coyle M, Dearden PK, Du Pasquier L, Duncan EJ, Ebert D, Eibner C, Erikson G, Evans PD, Extavour CG, Francisco L, Gabaldón T, Gillis WJ, Goodwin-Horn EA, Green JE, Griffiths-Jones S, Grimmelikhuijzen CJP, Gubbala S, Guigó R, Han Y, Hauser F, Havlak P, Hayden L, Helbing S, Holder M, Hui JHL, Hunn JP, Hunnekuhl VS, Jackson L, Javaid M, Jhangiani SN, Jiggins FM, Jones TE, Kaiser TS, Kalra D, Kenny NJ, Korchina V, Kovar CL, Kraus FB, Lapraz F, Lee SL, Lv J, Mandapat C, Manning G, Mariotti M, Mata R, Mathew T, Neumann T, Newsham I, Ngo DN, Ninova M, Okwuonu G, Ongeri F, Palmer WJ, Patil S, Patraquim P, Pham C, Pu L-L, Putman NH, Rabouille C, Ramos OM, Rhodes AC, Robertson HE, Robertson HM, Ronshaugen M, Rozas J, Saada N, Sánchez-Gracia A, Scherer SE, Schurko AM, Siggens KW, Simmons D, Stief A, Stolle E, Telford MJ, Tessmar-Raible K, Thornton R, Van der Zee M, von Haeseler A, Williams JM, Willis JH, Wu Y, Zou X, Lawson D, Muzny DM, Worley KC, Gibbs RA, Akam M, Richards S. 2014. The first myriapod genome sequence reveals conservative arthropod gene content and genome organisation in the centipede *Strigamia maritima*. *PLoS Biology* **12**:e1002005 DOI 10.1371/journal.pbio.1002005.

Clarke TH, Garb JE, Hayashi CY, Haney RA, Lancaster AK, Corbett S, Ayoub NA. 2014. Multi-tissue transcriptomics of the black widow spider reveals expansions, co-options, and functional processes of the silk gland gene toolkit. *BMC Genomics* **15**:365 DOI 10.1186/1471-2164-15-365.

Collins NM, Wells SM. 1987. *Invertebrates in need of special protection in Europe. Augier H. Nature & Environment Series No. 35*. Strasbourg: Council of Europe, pp. 162.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**:3674–3676 DOI 10.1093/bioinformatics/bti610.

Corzo G, Gilles N, Satake H, Villegas E, Dai L, Nakajima T, Haupt J. 2003. Distinct primary structures of the major peptide toxins from the venom of the spider *Macrothele gigas* that bind to sites 3 and 4 in the sodium channel. *FEBS Letters* **547**:43–50 DOI 10.1016/S0014-5793(03)00666-5.

Croset V, Rytz R, Cummins SF, Budd A, Brawand D, Kaessmann H, Gibson TJ, Benton R. 2010. Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction. *PLoS Genetics* **6**:e1001064 DOI 10.1371/journal.pgen.1001064.

Eddy SR. 2009. A new generation of homology search tools based on probabilistic inference. *Genome Informatics. International Conference on Genome Informatics* **23**:205–211.

Eisenberg E, Levanon EY. 2013. Human housekeeping genes, revisited. *Trends in Genetics* **29**:569–574 DOI 10.1016/j.tig.2013.05.010.

Foelix RF. 1970. Chemosensitive hairs in spiders. *Journal of Morphology* **132**:313–333 DOI 10.1002/jmor.1051320306.

Foelix RF, Chu-Wang IW. 1973. The morphology of spider sensilla. II. Chemoreceptors. *Tissue & Cell* **5**:461–478 DOI 10.1016/S0040-8166(73)80038-2.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**:3150–3152 DOI 10.1093/bioinformatics/bts565.

Gao L, Shan B, Chen J, Liu J, Song D, Zhu B. 2005. Effects of spider *Macrothele raven* venom on cell proliferation and cytotoxicity in HeLa cells. *Acta Pharmacologica Sinica* **26**:369–376 DOI 10.1111/j.1745-7254.2005.00052.x.

Grbić M, Van Leeuwen T, Clark RM, Rombauts S, Rouzé P, Grbić V, Osborne EJ, Dermauw W, Thi Ngoc PC, Ortego F, Hernández-Crespo P, Diaz I, Martinez M, Navajas M, Sucena É, Magalhães S, Nagy L, Pace RM, Djuranović S, Smagghe G, Iga M, Christiaens O, Veenstra JA, Ewer J, Villalobos RM, Hutter JL, Hudson SD, Velez M, Yi S V, Zeng J, Pires-daSilva A, Roch F, Cazaux M, Navarro M, Zhurov V, Acevedo G, Bjelica A, Fawcett JA, Bonnet E, Martens C, Baele G, Wissler L, Sanchez-Rodriguez A, Tirry L, Blais C, Demeestere K, Henz SR, Gregory TR, Mathieu J, Verdon L, Farinelli L, Schmutz J, Lindquist E, Feyereisen R, Van de Peer Y. 2011. The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. *Nature* **479**:487–492 DOI 10.1038/nature10640.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A. 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* **8**:1494–1512 DOI 10.1038/nprot.2013.084.

Harvey MS. 2002. Short-range endemism amongst the Australian fauna: some examples from non-marine environments. *Invertebrate Systematics* **16**:555–570 DOI 10.1071/IS02009.

Huang X. 1999. CAP3: a DNA sequence assembly program. *Genome Research* **9**:868–877 DOI 10.1101/gr.9.9.868.

Hung S-W, Wang T-L. 2004. Arachnid envenomation in Taiwan. *Ann Disaster Med* **3**:12–17.

Jiménez-Valverde A, Decae AE, Arnedo MA. 2011. Environmental suitability of new reported localities of the funnel-web spider *Macrothele calpeiana*: an assessment using potential distribution modelling with presence-only techniques. *Journal of Biogeography* **38**:1213–1223 DOI 10.1111/j.1365-2699.2010.02465.x.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**:1236–1240 DOI 10.1093/bioinformatics/btu031.

Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**:27–30 DOI 10.1093/nar/28.1.27.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**:772–780 DOI 10.1093/molbev/mst010.

Kornobis E, Cabellos L, Aguilar F, Frías-López C, Rozas J, Marco J, Zardoya R. 2015. TRUFA: a user-friendly web server for *de novo* RNA-seq analysis using cluster computing. *Evolutionary Bioinformatics* **11**:97–104 DOI 10.4137/EBO.S23873.

Kronestedt T. 1979. Study on chemosensitive hairs in wolf spiders (Araneae, Lycosidae) by scanning electron microscopy. *Zoologica Scripta* **8**:279–285 DOI 10.1111/j.1463-6409.1979.tb00639.x.

Lam KC, Mühlpfordt F, Vaquerizas JM, Raja SJ, Holz H, Luscombe NM, Manke T, Akhtar A. 2012. The NSL complex regulates housekeeping genes in *Drosophila*. *PLoS Genetics* **8**:e1002736 DOI 10.1371/journal.pgen.1002736.

Liu Z, Zhao Y, Li J, Xu S, Liu C, Zhu Y, Liang S. 2012. The venom of the spider *Macrothele raveni* induces apoptosis in the myelogenous leukemia K562 cell line. *Leukemia Research* **36**:1063–1066 DOI 10.1016/j.leukres.2012.02.025.

Mattila TM, Bechsgaard JS, Hansen TT, Schierup MH, Bilde T. 2012. Orthologous genes identified by transcriptome sequencing in the spider genus *Stegodyphus*. *BMC Genomics* **13**:70 DOI 10.1186/1471-2164-13-70.

Montagné N, de Fouchier A, Newcomb RD, Jacquin-Joly E. 2015. Advances in the identification and characterization of olfactory receptors in insects. *Progress in Molecular Biology and Translational Science* **130**:55–80 DOI 10.1016/bs.pmbts.2014.11.003.

Opatova V, Arnedo MA. 2014. From Gondwana to Europe: inferring the origins of Mediterranean *Macrothele* spiders (Araneae: Hexathelidae) and the limits of the family Hexathelidae. *Invertebrate Systematics* **28**:361–374 DOI 10.1071/IS14004.

Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**:1061–1067 DOI 10.1093/bioinformatics/btm071.

Parra G, Bradnam K, Ning Z, Keane T, Korf I. 2009. Assessing the gene space in draft genomes. *Nucleic Acids Research* **37**:289–297 DOI 10.1093/nar/gkn916.

Patel RK, Jain M. 2012. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS ONE* **7**:e30619 DOI 10.1371/journal.pone.0030619.

Pelosi P, Iovinella I, Felicioli A, Dani FR. 2014. Soluble proteins of chemical communication: an overview across arthropods. *Frontiers in Physiology* **5**:Article 320 DOI 10.3389/fphys.2014.00320.

Platnick NI. 2006. The world spider catalog, V6.5 by N. I. Platnick. AMNH. *Available at https://research.amnh.org/iz/spiders/catalog_15.0/index.html*.

Posnien N, Zeng V, Schwager EE, Pechmann M, Hilbrant M, Keefe JD, Damen WGM, Prpic N-M, McGregor AP, Extavour CG. 2014. A comprehensive reference transcriptome resource for the common house spider *Parasteatoda tepidariorum*. *PLoS ONE* **9**:e104885 DOI 10.1371/journal.pone.0104885.

Prosdocimi F, Bittencourt D, da Silva FR, Kirst M, Motta PC, Rech EL. 2011. Spinning gland transcriptomics from two main clades of spiders (order: Araneae)–insights on their molecular, anatomical and behavioral evolution. *PLoS ONE* **6**:e21634 DOI 10.1371/journal.pone.0021634.

Sánchez-Gracia A, Vieira FG, Rozas J. 2009. Molecular evolution of the major chemosensory gene families in insects. *Heredity* **103**:208–216 DOI 10.1038/hdy.2009.55.

Sanggaard KW, Bechsgaard JS, Fang X, Duan J, Dyrlund TF, Gupta V, Jiang X, Cheng L, Fan D, Feng Y, Han L, Huang Z, Wu Z, Liao L, Settepani V, Thøgersen IB, Vanthournout B, Wang T, Zhu Y, Funch P, Enghild JJ, Schauser L, Andersen SU, Villesen P, Schierup MH, Bilde T, Wang J. 2014. Spider genomes provide insight into composition and evolution of venom and silk. *Nature Communications* **5**:Article 3765 DOI 10.1038/ncomms4765.

Satake H, Villegas E, Oshiro N, Terada K, Shinada T, Corzo G. 2004. Rapid and efficient identification of cysteine-rich peptides by random screening of a venom gland cDNA library from the hexathelid spider *Macrothele gigas*. *Toxicon* **44**:149–156 DOI 10.1016/j.toxicon.2004.05.012.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**:1312–1313 DOI 10.1093/bioinformatics/btu033.

Vieira FG, Rozas J. 2011. Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: origin and evolutionary history of the chemosensory system. *Genome Biology and Evolution* **3**:476–490 DOI 10.1093/gbe/evr033.

Yamaji N, Little MJ, Nishio H, Billen B, Villegas E, Nishiuchi Y, Tytgat J, Nicholson GM, Corzo G. 2009. Synthesis, solution structure, and phylum selectivity of a spider delta-toxin that slows inactivation of specific voltage-gated sodium channel subtypes. *The Journal of Biological Chemistry* **284**:24568–24582 DOI 10.1074/jbc.M109.030841.

Zeng X-Z, Xiao Q-B, Liang S-P. 2003. Purification and characterization of raventoxin-I and raventoxin-III, two neurotoxic peptides from the venom of the spider *Macrothele raveni*. *Toxicon* **41**:651–656 DOI 10.1016/S0041-0101(02)00361-6.

Zhang S, Koh TH, Seah WK, Lai YH, Elgar MA, Li D. 2012. A novel property of spider silk: chemical defence against ants. *Proceedings Biological Sciences of the Royal Society* **279**:1824–1830 DOI 10.1098/rspb.2011.2193.

B

Financiación

Esta tesis doctoral ha estado financiada por el Ministerio de Economía y Competitividad (CGL2013-45211 y CGL2016-75255), la Comissió Interdepartamental de Recerca I Innovació Tecnològica (2014SGR-1055 y 2017SGR1287) y parcialmente financiada por ICREA Academia (Generalitat de Catalunya). Durante el periodo pre-doctoral, Joel Vizueta Moraga ha disfrutado de una Beca de Formación de Personal Investigador (FPI; BES-2014-068437) del Ministerio de Economía y Competitividad.